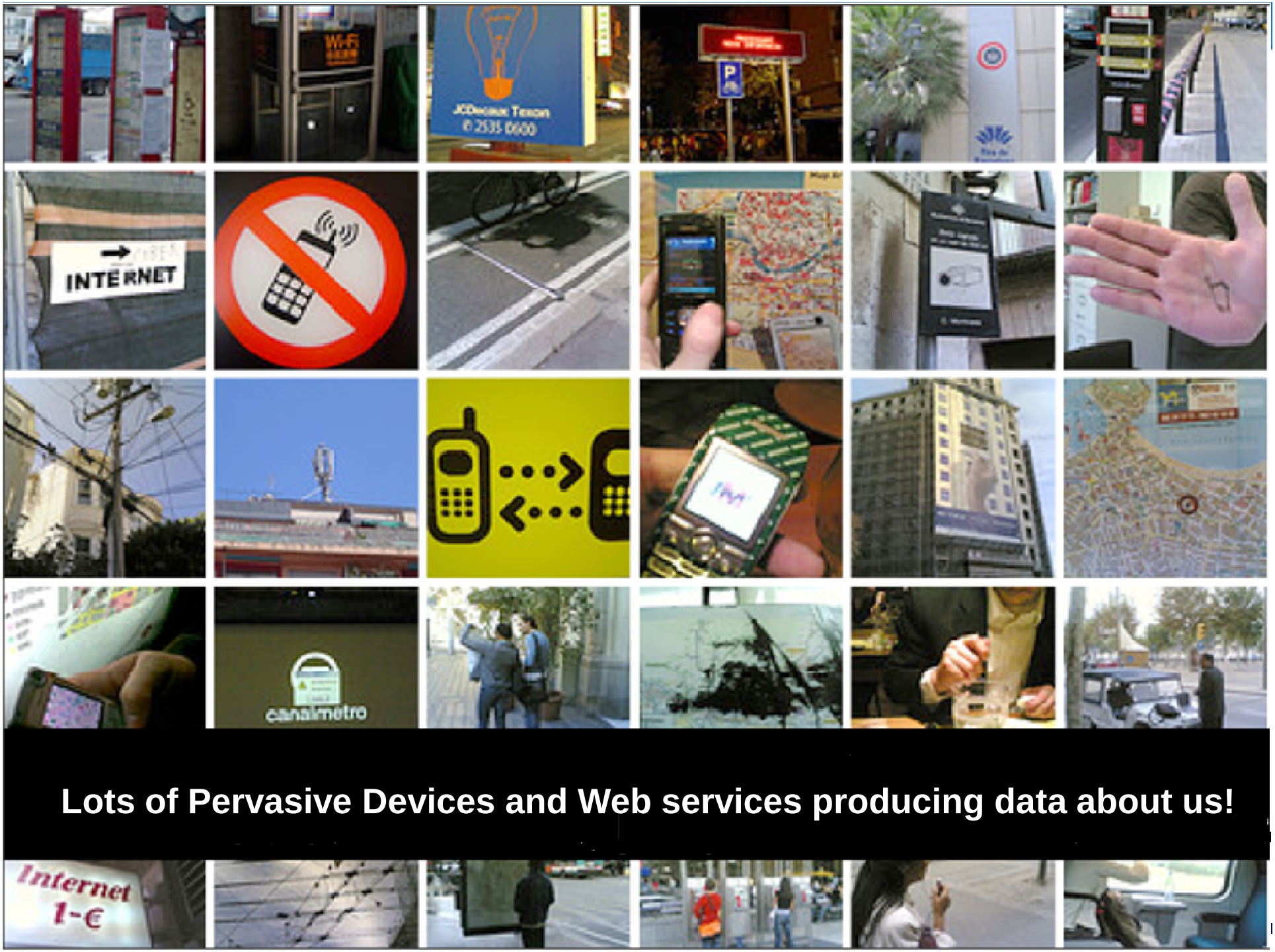
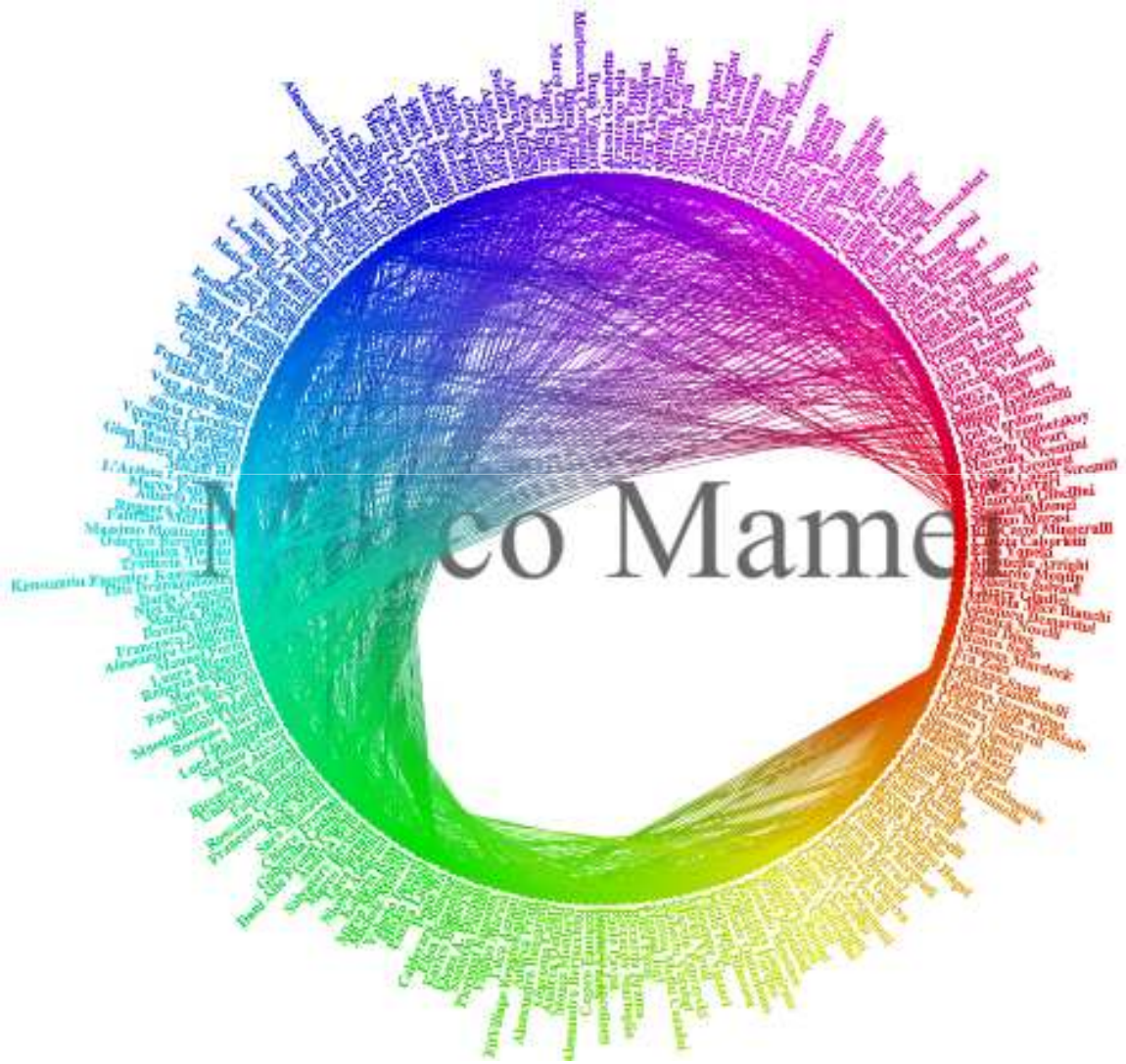
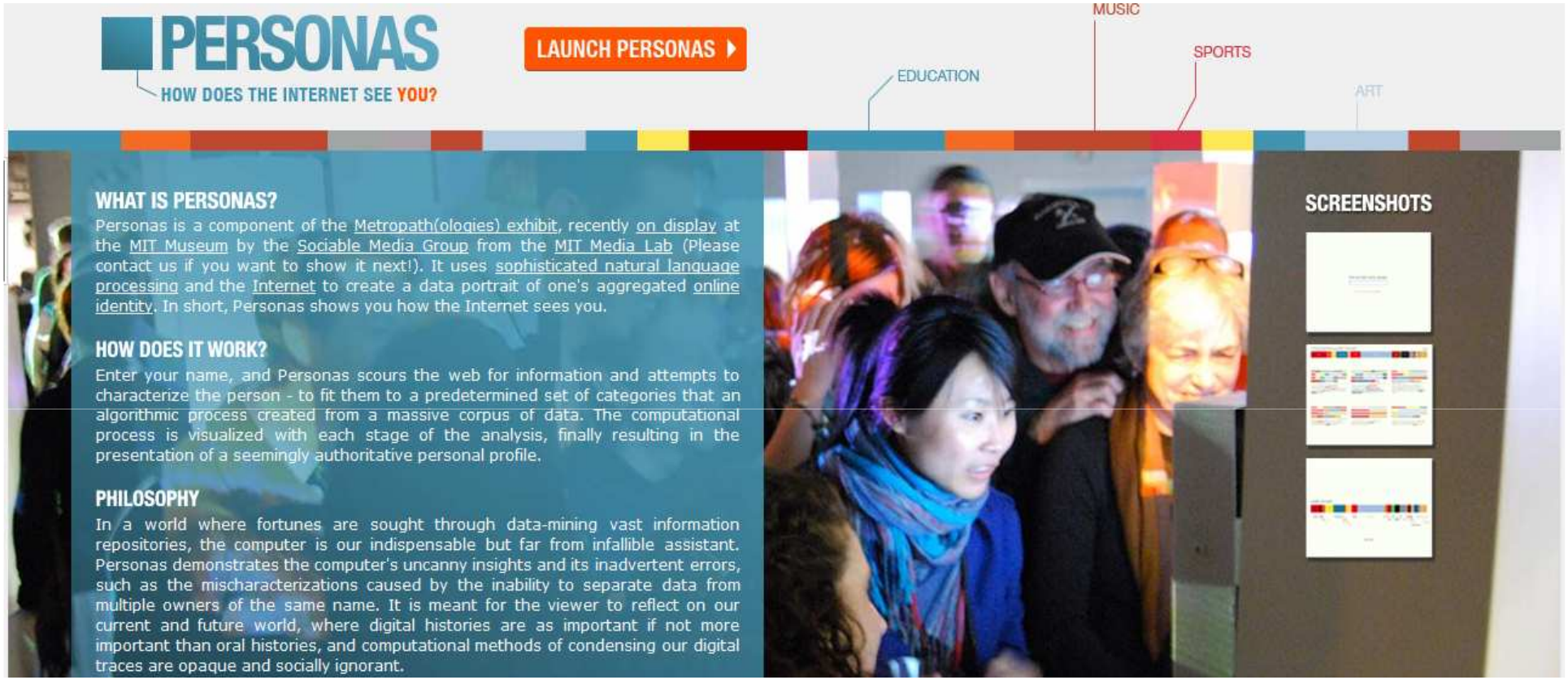
Knowledge Extraction from Mobility Data

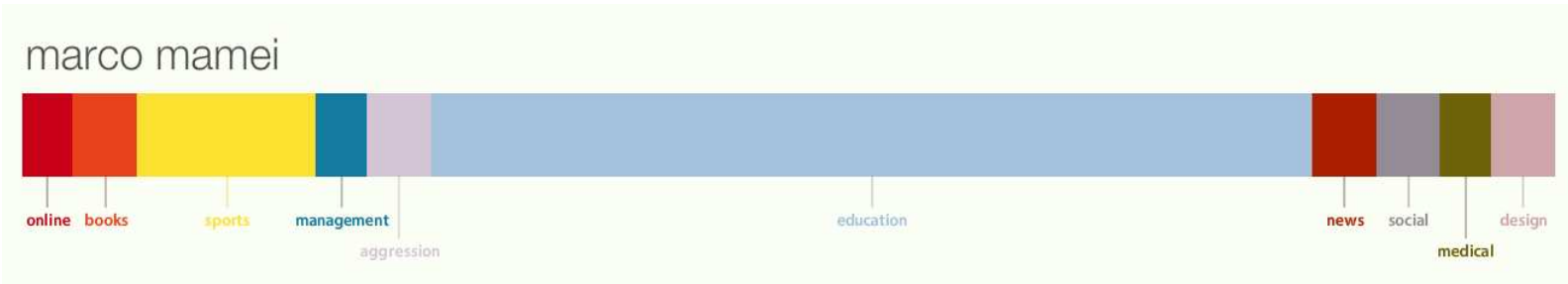
Marco Mamei



Lots of Pervasive Devices and Web services producing data about us!


Marco Mamei


PERSONAS
HOW DOES THE INTERNET SEE YOU?
LAUNCH PERSONAS
MUSIC
SPORTS
EDUCATION
ART
WHAT IS PERSONAS?
Personas is a component of the Metropath(ologies) exhibit, recently on display at the MIT Museum by the Sociable Media Group from the MIT Media Lab (Please contact us if you want to show it next!). It uses sophisticated natural language processing and the Internet to create a data portrait of one's aggregated online identity. In short, Personas shows you how the Internet sees you.
HOW DOES IT WORK?
Enter your name, and Personas scours the web for information and attempts to characterize the person - to fit them to a predetermined set of categories that an algorithmic process created from a massive corpus of data. The computational process is visualized with each stage of the analysis, finally resulting in the presentation of a seemingly authoritative personal profile.
PHILOSOPHY
In a world where fortunes are sought through data-mining vast information repositories, the computer is our indispensable but far from infallible assistant. Personas demonstrates the computer's uncanny insights and its inadvertent errors, such as the mischaracterizations caused by the inability to separate data from multiple owners of the same name. It is meant for the viewer to reflect on our current and future world, where digital histories are as important if not more important than oral histories, and computational methods of condensing our digital traces are opaque and socially ignorant.
SCREENSHOTS


marco mamei
online
books
sports
management
aggression
education
news
social
medical
design

File View Tools Help

Timeline Favourites

NOKIA Lifeblog

17/03/2004

Tuesday 16


07:53 Bike

Ooo. Sounds so lovely.
18:09 To: Mary


22:56 To: David


10:04 Tom working

In your absence we have decided the following:
1) No curtains 2) Layout of the big space as mirror image of planned.
23:08 Text Note


15:27 My office

Wednesday 17


11:39 Museum


19:21 To: Tom

I feel terrible for not being there. You're my hero! Do you want to come bowling with me and a few friends tonight? 8:30 pm at University pl between 12th and 13th st.
11:39 From: Jane


21:37 Mary writing


14:11 From: David


23:09 Open space

Thursday 18


06:53 image024.jpg

Believe it or not, we've started exercising. Jump roping at the terrace every morning! Next picture will be of my athletic body. Email me some work when you get back.
09:46 From: Jane

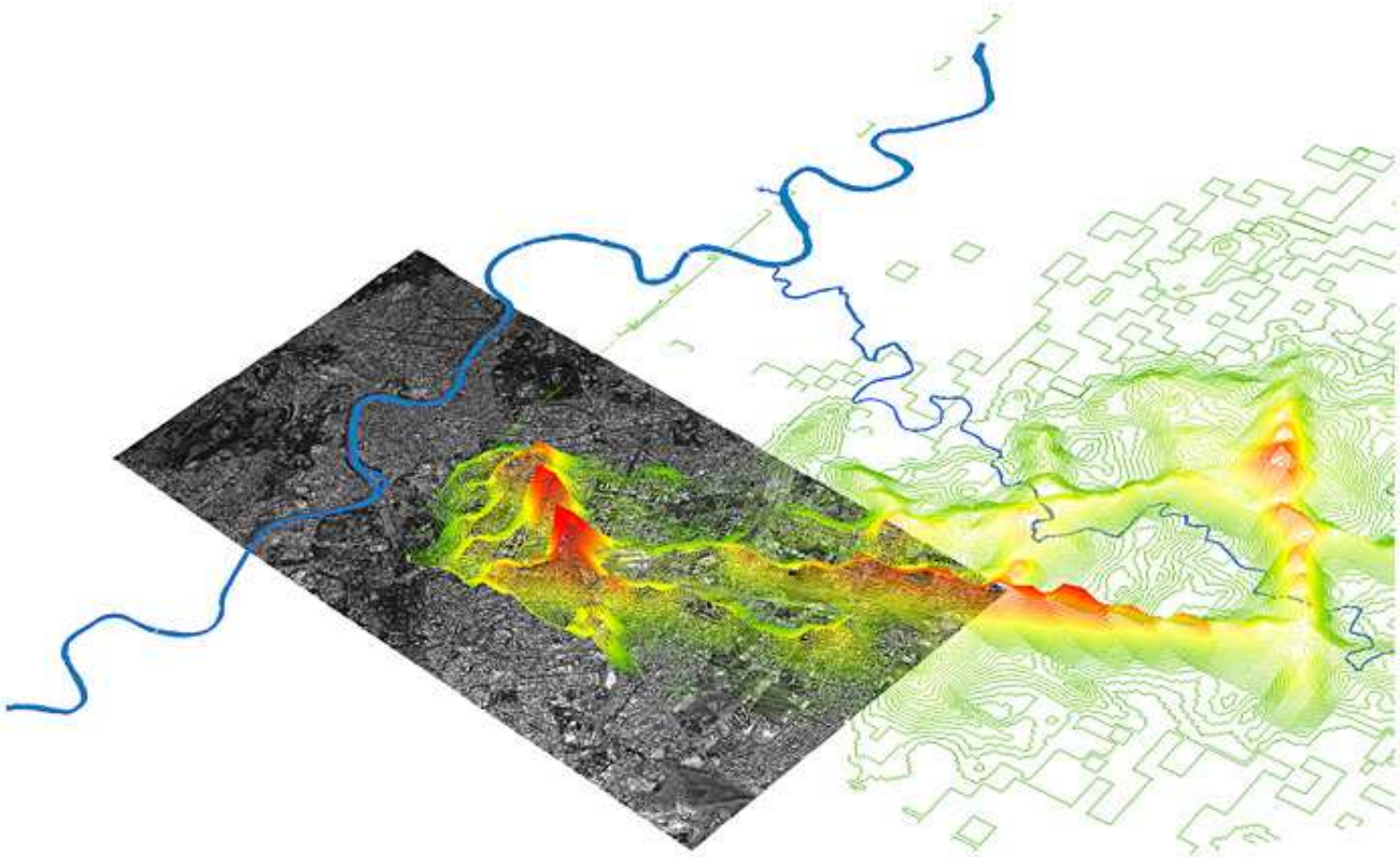

10:02 At work

Filters...

Favourites
Always on your phone


DINING
Selected findings from meals at 81 restaurants, 13 homes and four events
141
MOMOFUKU SSAM
17
FLYING FISH
SOUS VIDE SALMON
HOME COOKED
AMERICAN
FUGU, FROG LEGS & GOAT
ONE
MIRACLE FRUIT
DRINKING
Regarding beverages consumed at 73 restaurants, 52 bars, 25 homes, seven events and two offices.
573
SEVEN
ASSORTED
STELLA
SIERRA NEVADA ESB
FIG, MINT & TEQUILA
STELLA ARTOIS
408
2.9
ONE

Copenaghen Wheel



Outline

- Mobility Data and Applications
- Long-term mobility data
  - The whereabouts diary
  - Routine extraction from data
- Short-term mobility data
  - POI discovery from Flickr photo stream
  - Sport city dynamics from Nokia Sport Tracker
- Future directions

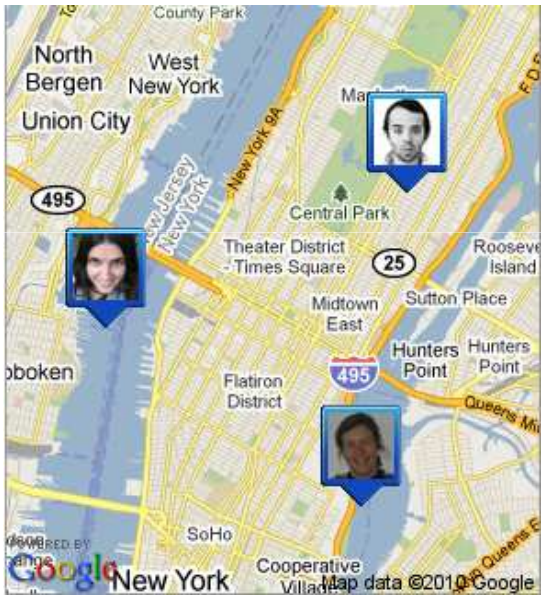
Mobility Data

- Mobility data is one of the first example of data from pervasive technology going mainstream.
- It is a first link between the Web and the physical world.
- The number and availability of whereabouts data is rapidly increasing…
  - Google Latitude
  - Yahoo Fire Eagle / Friends on Fire
  - FourSquares
  - Facebook Places
  - Gowalla
  - Geotagged photo (Flickr, Picasa)
  - Geotagged tweets

Applications

- The number of applications that can take advantage of such data is huge

- Maps and navigation
- Location-based search
- Location-based personalized seraches
- Location-based social networks



- Novel application rely on the fact that mobility data is a mean to gather information about users and their environment

User-centerd Applications

- **Pervasive Advertisement.** An application could show commercials to the user that are personalized on the basis of the diary.
- **Tourists recommendatons.** You like museums, the application recommends other similar places.
- **Personalized Navigation.** navigation routes with the goal of reducing route complexity and cognitive burden.
- **Life Logging, Life Blogging**

Environment-centered Applications

- **Identify places and POI.** "*Which are the most crowded pubs on Saturday night?*", "*Which are the restaurants visited by people living in my neighborhood?*", etc. The results can be used to retrieve and recommend Web content.

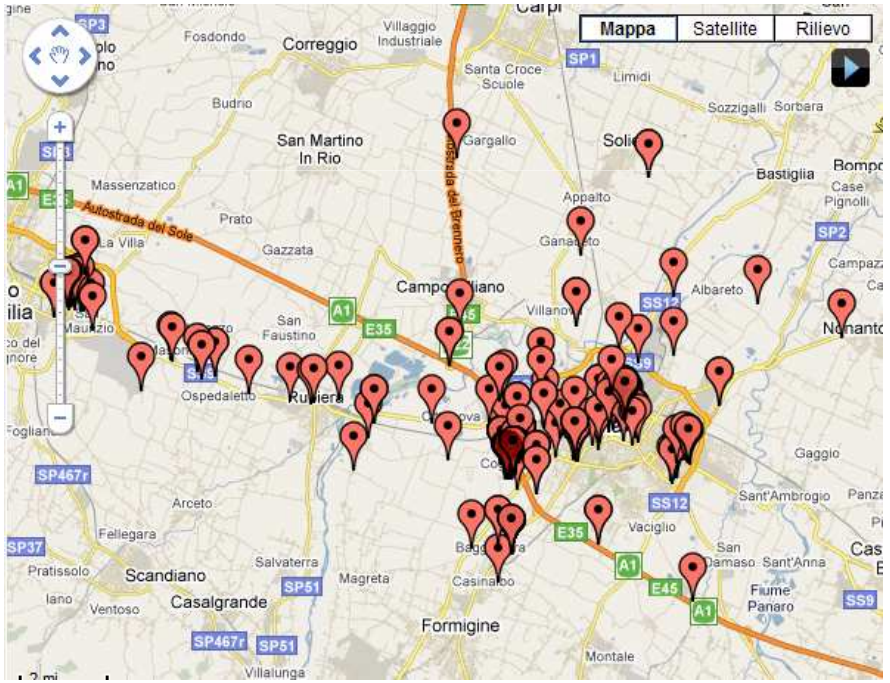
- **Identify events.** If a large number of people visit a specific location in Barcelona, say Camp Nou, on the same day, we may infer that there is an important event, such as a concert or a soccer game, happening at that location.

- **Urban Planning.** Mobility data may be used to inform how businesses or infrastructure are distributed across the city, so as to foster their placement (and opening time) where they are most required and would be most useful.

- **Disaster recovery scenarios**, the actual distribution of people at the time of the disaster (*e.g.*, earthquake) could be a critical asset to organize a contingency plan and prioritize resources. An analysis and prediction of where people are in the city at certain times of the day and year can be combined with locations of hospitals, doctors and transportation.

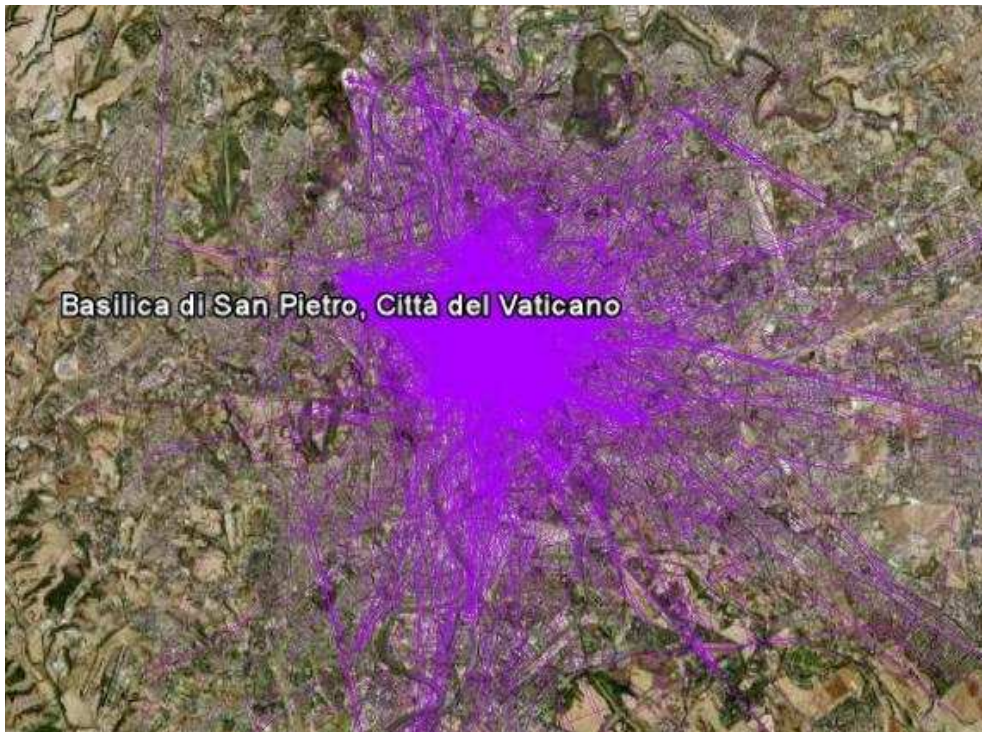
Theoretical Challenges

- Making sense of data
  - How to code personalized advertisement?



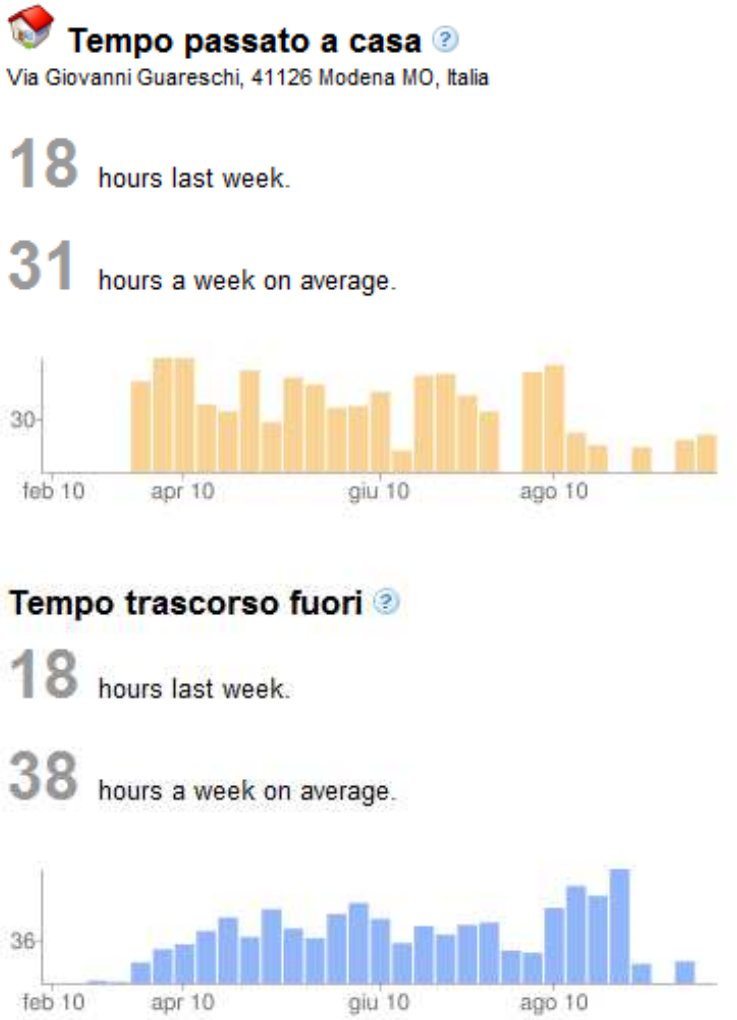
Theoretical Challenges

- Making sense of data
  - What are the POI?



Theoretical Challenges

- So as to provide usable knowledge to applications



Practical Challenges

- Get access to dataset
- Get access to ground truth information

- **Long term mobility**. Tracking a user 24by7
  - Difficult to get large dataset
  - Strong privacy issues
  - Geared toward user-centered applications
  - *Example.* Google Latitude Data.
- **Short term mobility**. Trackng a user during specific activities (e.g., taking a picture)
  - Easier to get large dataset (but it is never large enough)
  - Hard to get groundtruth data
  - Geared toward environment-centered applications
  - *Example.* Flickr photos

Privacy

	LOCATIONS	
TIME	MEGAN	ROBER
11:00 AM	HOME	
12:30 PM	EASTVIEW ADULT TOY STORE	HOME
1:30 PM	HOME	
2:00 PM	LAKETOWN SEX TOY SHOP	SCHOO
2:30 PM	HOME	
3:00 PM	FRY'S ELECTRONICS	
3:30 PM	ED'S POWER TOOL EMPORIUM	SUBWA
4:00 PM	HOME	
4:10 PM	HOSPITAL BURN WARD	

Four Exemplary Researches

To show possible approache to tackle the above challenges
both in the long and in the short scale

1. The Whereabouts Diary

Long-term tracking of GPS traces

What is it?

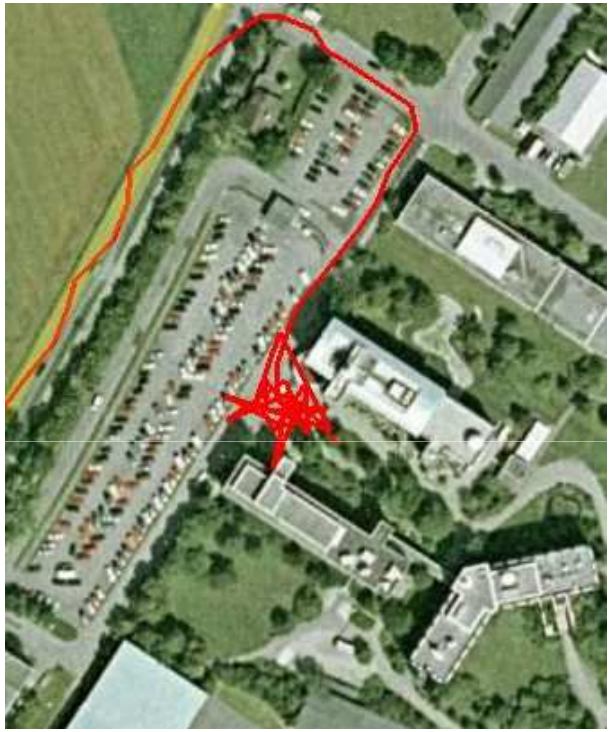
- The whereabouts diary is an application, running on a GPS-equipped handheld device that records the list of relevant places visited by the user. The diary runs autonomously without requiring user's interactions and is able to classify *semantically* the places being visited in an unsupervised way.

- The places we go can reveal something about us, and can be used as a surrogate or a complement to form a better user profile.
  - For example, a matchmaking application could infer that two persons are compatible given the fact that they visit almost the same places.
  - if the places are tagged semantically (e.g., work, home, pub, etc.) the application could infer more advanced relationships among the persons. For example, two persons visiting the same "work" place could be marked as colleagues, while persons visiting the same "home" place could be marked as relatives.

Creating the Diary

- The construction of the diary is an incremental process
- Starting from the log of the GPS readings (or of other kind of localization devices), it is possible to run segmentation and clustering algorithms to infer the places where the user has been

Diary based on GPS coordinates



Longitude	Latitude	Time
11°16'43.17"E	48° 5'11.75"N	Sept. 20, 2010, 8:35am-10:45am
…	…	…

Diary based on addresses

- Using inverse geocoding services it is possible to identify the addresses associated to the identified places.

Address	Time
H-1021 Budapest, Pálos utca 2, Hungary	**Sept. 20, 2010, 8:35am-10:45am**
…	…

- In general, because of errors in GPS readings multiple addresses are retrieved….

Diary based on places

- The diary can look for a particular address in yellow and white pages services to identify what is in a particular address.

Place	Time
Europa Hotels & Congress Center	**Sept. 20, 2010, 8:35am-10:45am**
…	…

- Moreover, the diary can mine the Web looking for what is happening in that place at that time.

Place	Time
Perada Assyst Summer School 2010	**Sept. 20, 2010, 8:35am-10:45am**
…	…

Diary based on personalized places

- If the user activities are profiled in some way (e.g., the diary may know a priori that the user tends to stay at home at night), then the diary application can give labels to places by looking at the temporal patterns in which places are visited. For example, the place most visited at night during weekdays can be meaningfully labeled as "Home".

Place	Time
Working place	**Sept. 20, 2007, 8:35am-10:45am**
…	…

- In its final form the diary represents a powerful source of context information allowing to extrapolate user's habits, preferences and routine behavior.

This is how the Whereabouts Diary should work….

Let’s see our implementation…

Diary based on GPS coordinates

- the GPS signal is lost for at least T seconds and it is re-acquired later on at a distance of less than L meters from where it was lost. This reflects the situation in which a user enters a building and leaves it after some time.
  - Some empirical evaluations let us to set T = 20 minutes, L = 20 meters.
  - The constraint on time is important to wash out GPS signal glitches,
  - the constraint on space is useful to avoid those situation in which the GPS has been shut down and the user moves away.

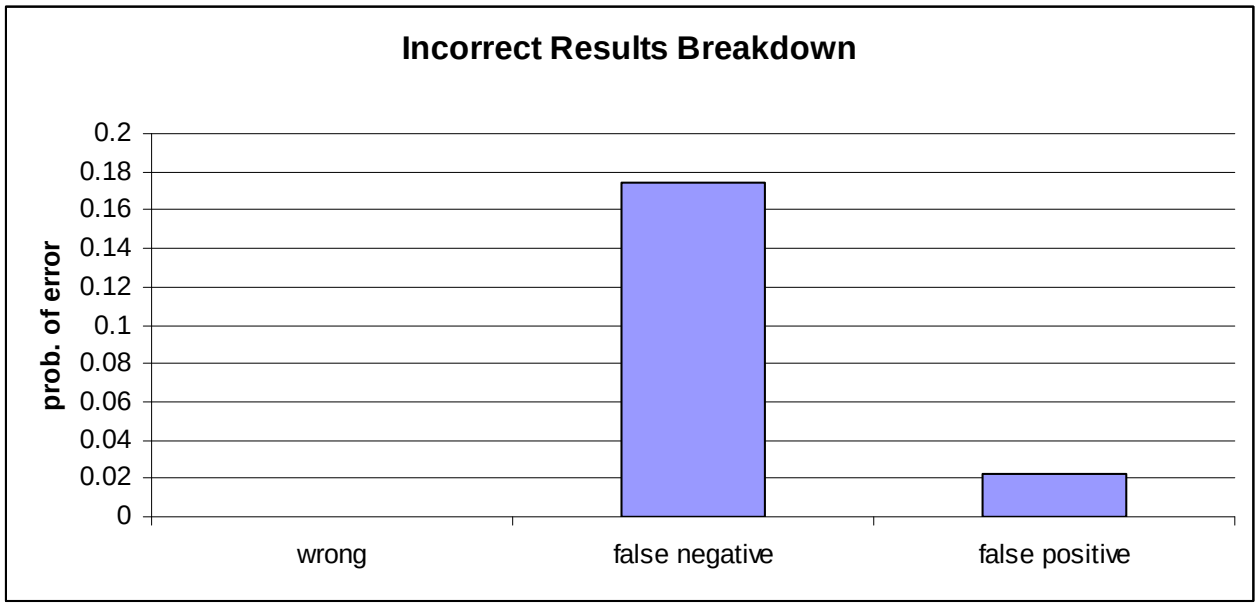
- The GPS readings over a time window of W seconds are clustered within a radius of R meters from each other. This reflects the situation in which the user stays for a long time in a place like a park or a square.
  - Some empirical evaluations let us to set W = 20 minutes, R = 100 meters.

Experiments Set up

- We collected our own GPS traces for 3 weeks as we went about our normal lives.
  - Each member carried a PDA connected with a Bluetooth GPS reader and running the Whereabouts Diary J2ME application.
  - GPS signal has been acquired at 0.1Hz and processed on the fly by the handheld device. Overall, 25 places were identified as relevant.
  - Ground-truth information about the places where we have been, were recorded.

GPS Performance

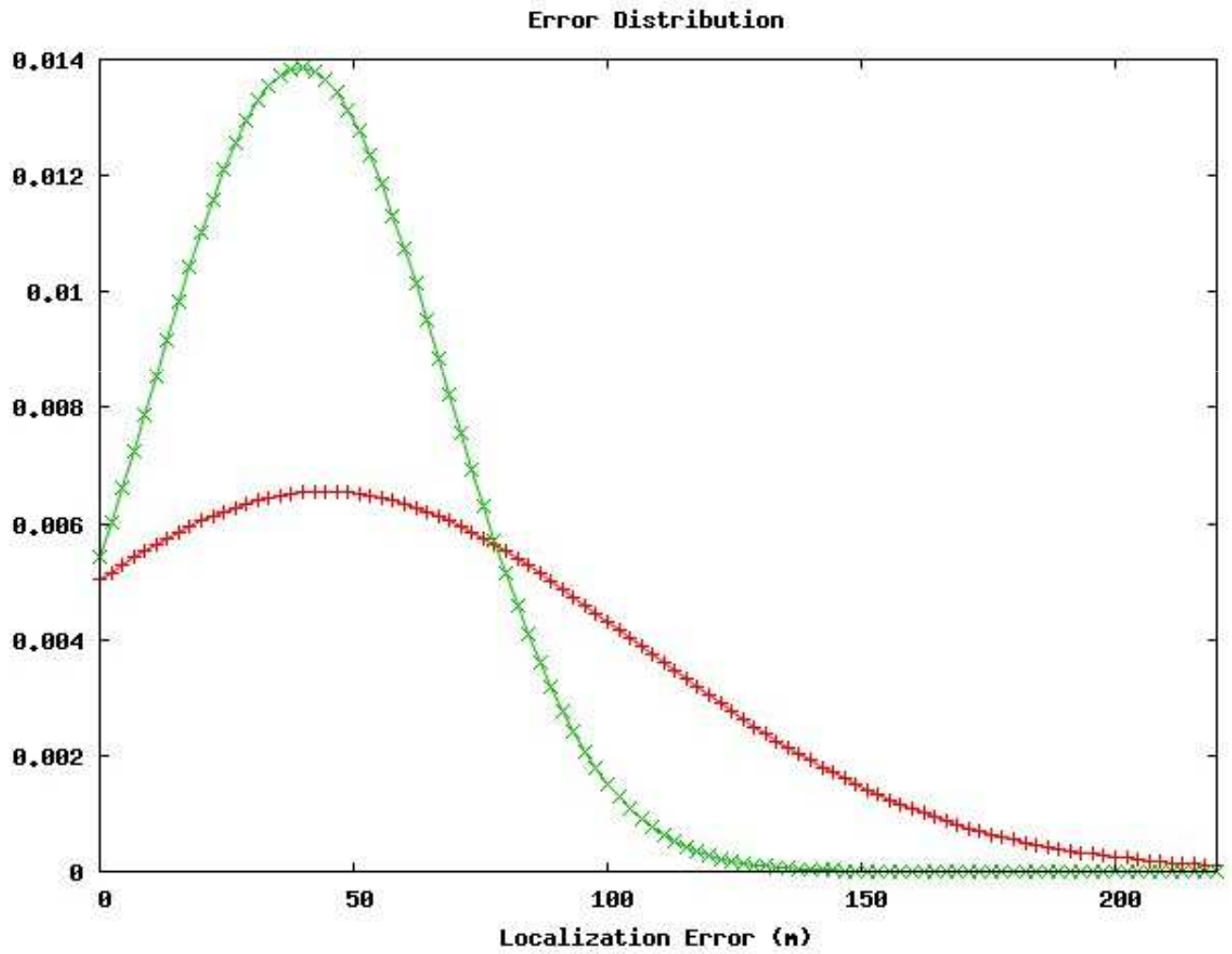
- The algorithm is correct in 84.7% of the cases (detected place is close (< 20 m) to the ground-truth data.



- *wrong*: the user is in a place, but the diary reports he is in a different place
- *false negative*: the user is in a place, but the diary reports he is moving
- *false positive*: the user is moving, but the diary reports he is in a place.

- The high-percentage of false negative results is due to the fact sometimes the GPS takes a long time before acquiring the signal. Thus, it can happen that a user leaves a building, and the trace of the GPS is acquired only when he is already far away

GPS Performances

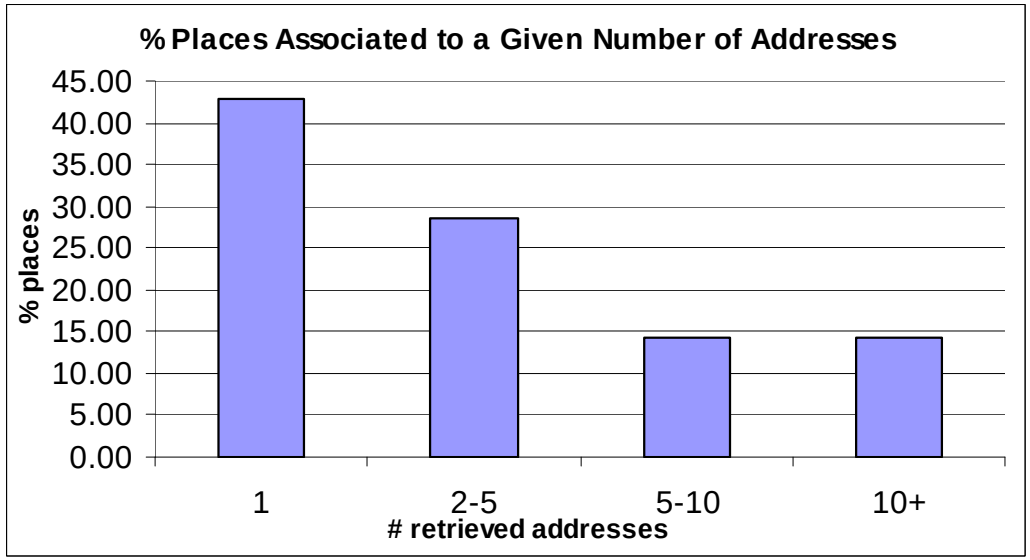


Diary based on addresses

- We developed a "reverse" geocoding for our region, on the basis of maps available from a commercial navigator software.
- Street numbers are evenly spread on the street length

- The coordinates are mapped to the closer map entry (i.e., address) being available.

Reverse Geocoding Performance

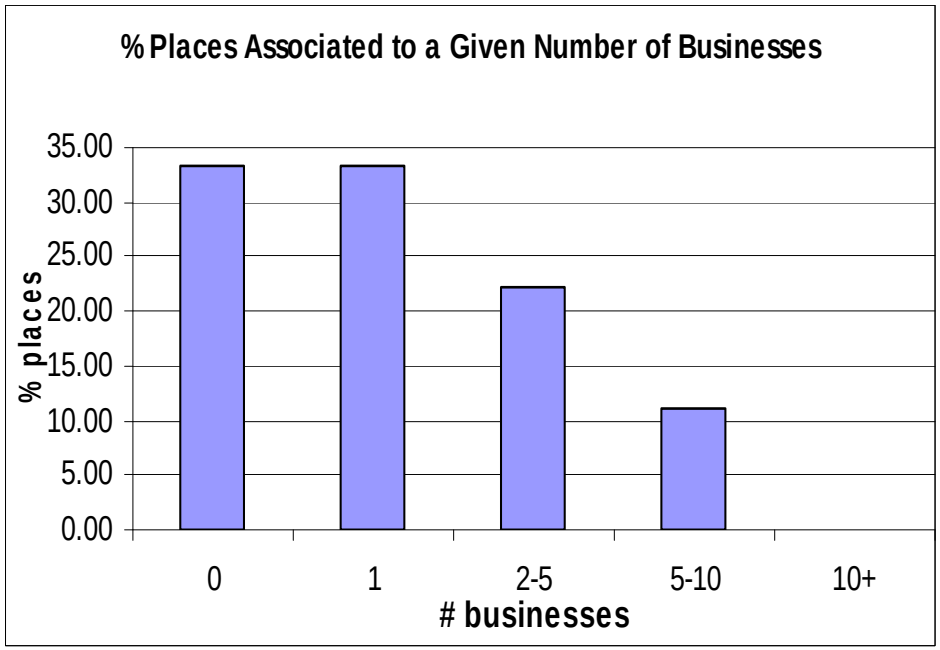


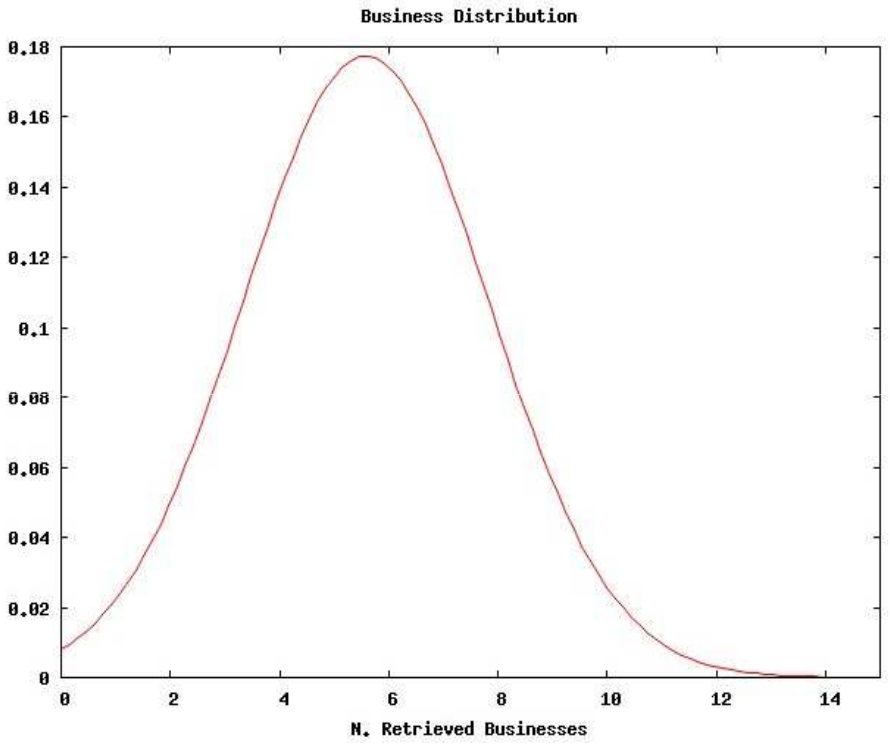
- The address of almost half of the places can be retrieved uniquely (this is the case of large buildings – like the departments of our university).
- Some places produce more than 10 associated addresses. This is the case of small buildings in the center of the city
- NOTE. Those distributions are based on the 25 identified places, thus they are not very stable…

Diary based on Places

- We screen-scraped information coming from a widely used online white-pages service (*www.paginebianche.it*) in our region allowing to query for who is at a given address.
- Each geocoded address belonging to a given place (as provided by the previous step) is looked up in the white-pages and the corresponding business is retrieved.

Business Search Performance

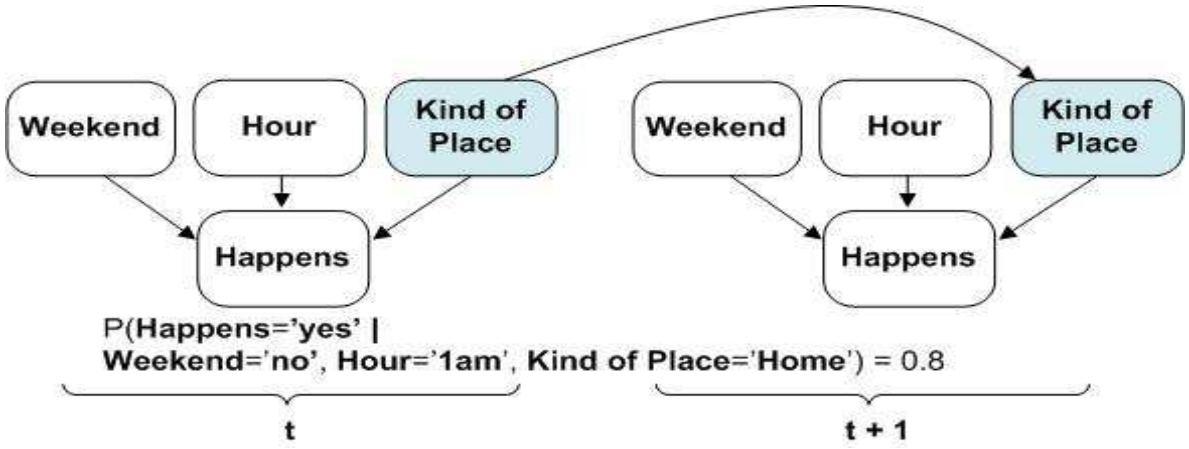




- The actual place can be retrieved in only 40% of the cases. Moreover, the number of businesses being retrieved is almost independent of whether the correct place has been found or not. This is either due to localization or white-pages errors

Diary based on Personalized Places

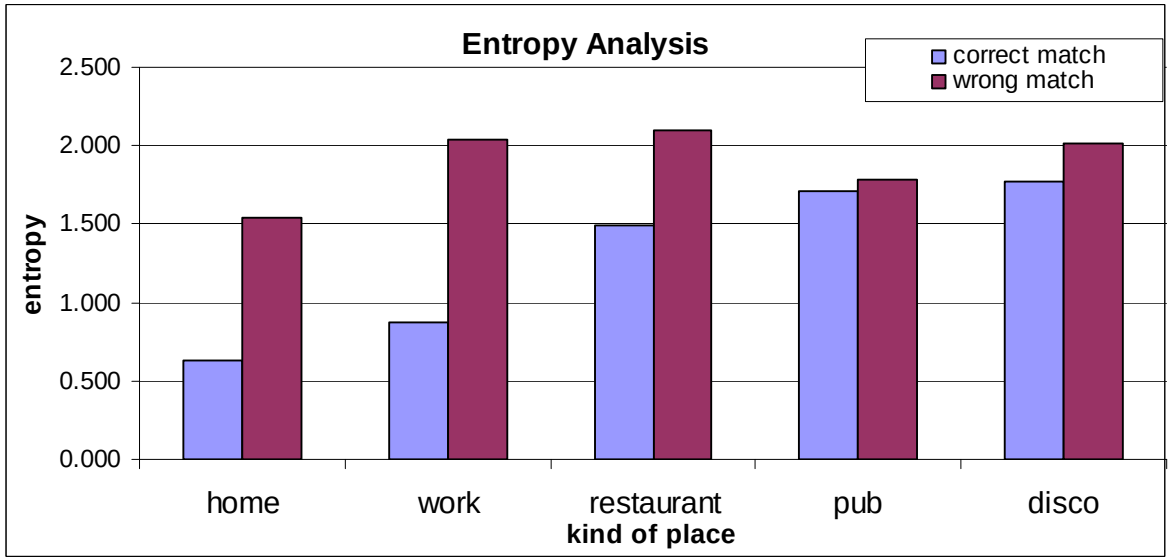
- For each place being identified, the diary creates a Bayesian network to analyze the temporal pattern in which the place has been visited by the user.



Weekend = false, Kind of Place = home									
time	**11pm-6am**	**7am**	**8am**	**9am-1pm**	**2pm-5pm**	**6pm-7pm**	**8pm**	**9pm**	**10pm**
P(happens) = true	0.8	0.6	0.4	0.2	0.2	0.4	0.5	0.6	0.7

Performance of the Bayesian Networks

- Overall, our approach classifies the places correctly in 64% of the cases.
- In order to better analyze the results we tried to assess the confidence of the diary in its own classification – most probable estimate (MPE). To this end, we compute the information entropy of the resulting distributions.
  - The lower the entropy, the more the system is confident about the MPE (i.e., the distribution peaks on the MPE value).



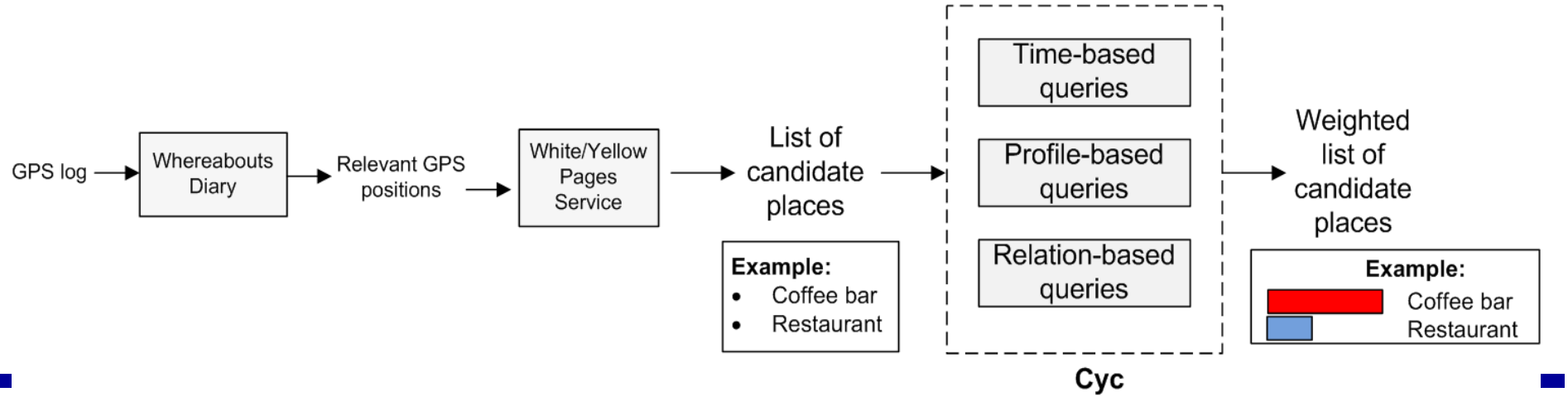
Discussion

- In the end, accuracy will be the key measure in which the diary will be evaluated. If the diary is wrong, the applications that use it risk being rendered useless.
  - Other kind of sensing devices and algorithms could be employed to extract more information about the place (e.g., credit card transaction record). Moreover, some GPS clustering techniques that have been used in some recent works could improve the performance of our implementation.
- It is important to evaluate the diary on real applications to see if its accuracy is enough to effectively support that application.

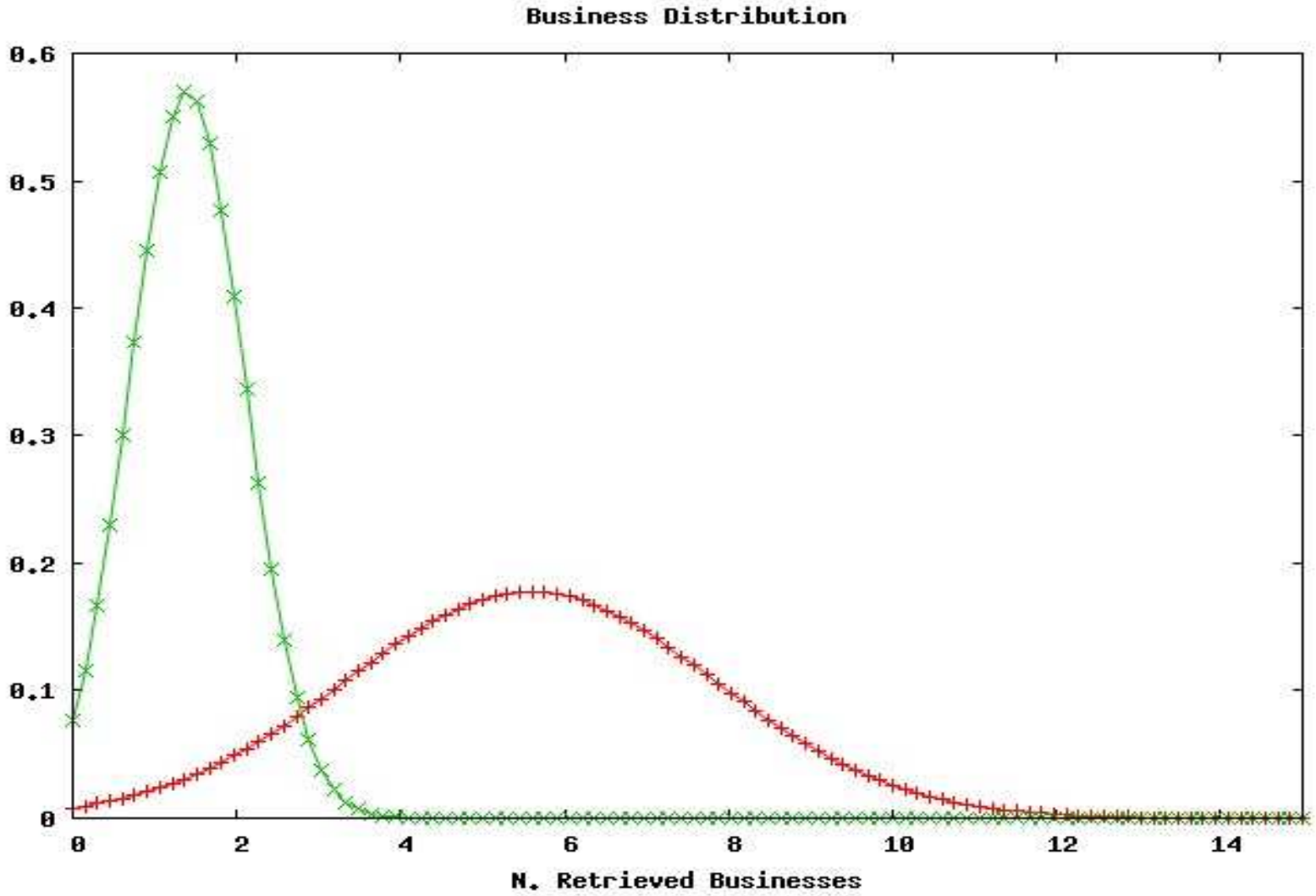
Integration with CYC Commonsense

- Commonsense data could be exploited to effectively discriminate among several candidate places. For example, if a person went to a restaurant at noon, it is very unlikely that will go to another restaurant at 2pm.

- **The CYC Knowledge Base (KB)** contains contains over a million human-defined assertions, rules or common sense ideas. These are formulated in the language CycL, which is based on predicate calculus.
- **The Inference Engine** allows to query the KB. It performs general logical deduction by using best-first search using proprietary heuristics



CYC Result
(preliminary)



2. Classification of Whereabouts Patterns from Large-Scale Mobility Data

Long-term tracking of GSM traces

Beyond the diary

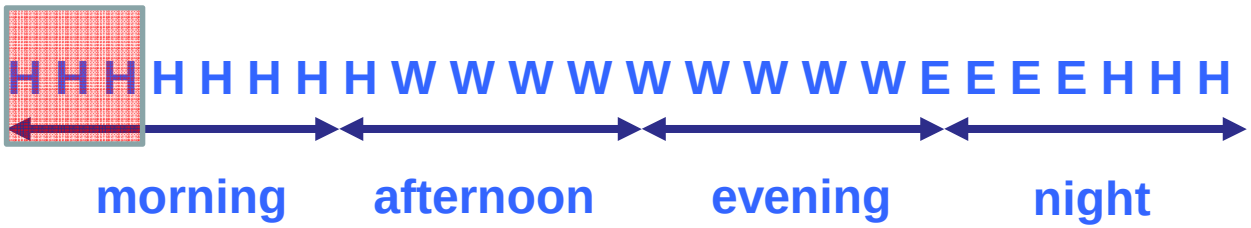
- Even in the most complete form, the diary represents user's daily life in a rahter episodic way

home	**Sept. 20, 2007, 00:35am-08:45am**
work	**Sept. 20, 2007, 09:35am-06:45pm**
home	**Sept. 20, 2007, 09:35pm-11:45pm**

- It would be interesting to identify routine and recrruent behaviors from such a log.
- Describe the above day as "day at work and pub with friends afterwards"

Routine Extraction… LDA

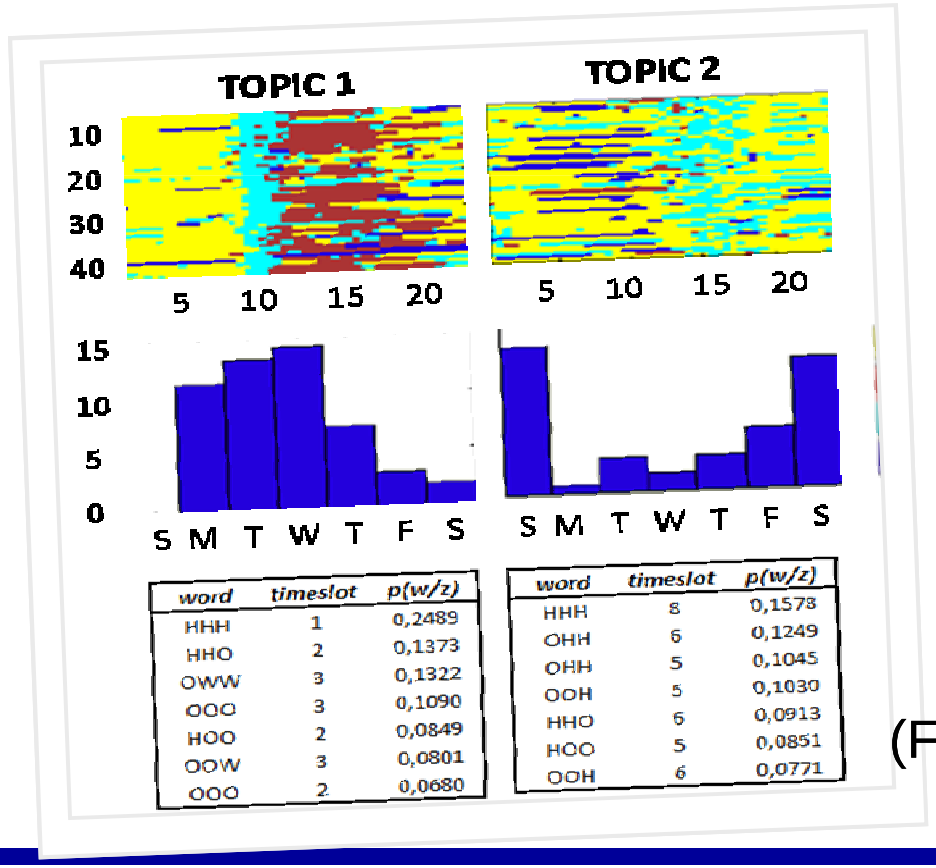
Place	Time
home	Sept. 20, 2007, 00:35am-08:45am
work	Sept. 20, 2007, 09:35am-06:45pm
pub	Sept. 20, 2007, 07:35pm-08:45am
home	Sept. 20, 2007, 09:35pm-11:45pm
…	…



HHH1, HHH1, …. HHW2, WWW2,…

LDA

- Probabilistic model clustering words (w) in topics (z).
- Words like HHH1, WWW2, WWW3, HHH4 will be clustered together in a topic Z expressing "normal working routine"



word	timeslot	p(w/z)
HHH	1	0,2489
HHO	2	0,1373
OWW	3	0,1322
OOO	3	0,1090
HOO	2	0,0849
OOW	3	0,0801
OOO	2	0,0680

word	timeslot	p(w/z)
HHH	8	0,1573
OHH	6	0,1249
OHH	5	0,1045
OOH	5	0,1030
HHO	6	0,0913
HOO	5	0,0851
OOH	6	0,0771

(Farrahi et al., 2009)

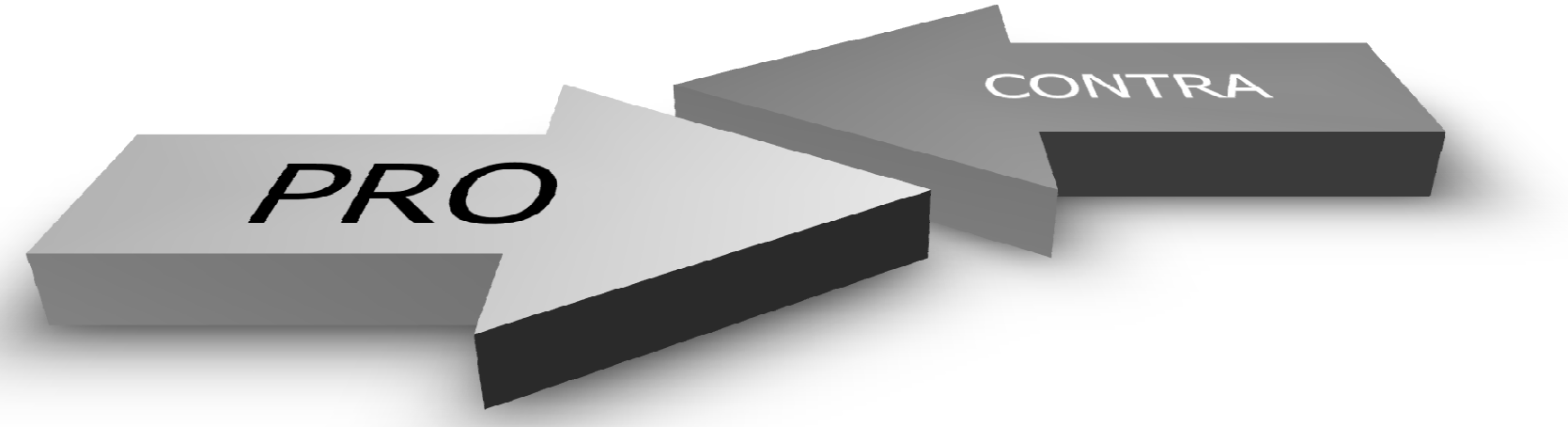
Problem Identification

PRO

- Topical pattern analysis
- Summarization
- Subtopic discovery

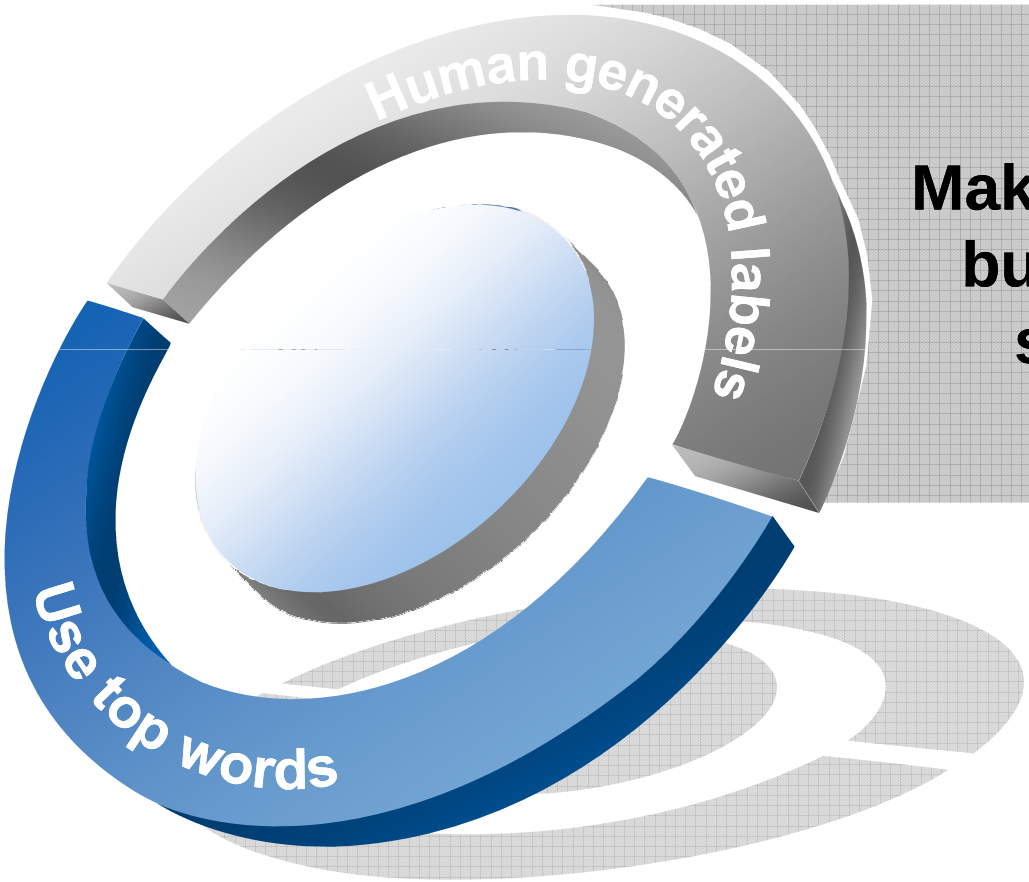
CONTRA

- Predefined number of topics
- Hard to interpret



Problem Identification

Automatic but hard to make sense

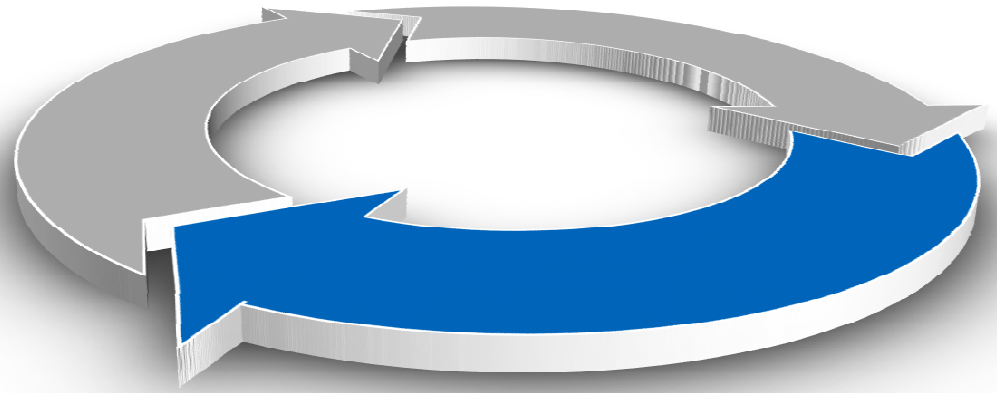


Make sense, but cannot scale up.

Research questions

can we identify patterns from mobility data?

can we automatically generate understandable labels for topics?



can we automatically attach labels to such behavioral patterns?

Applications of labeling patterns

create an entry in the user blog

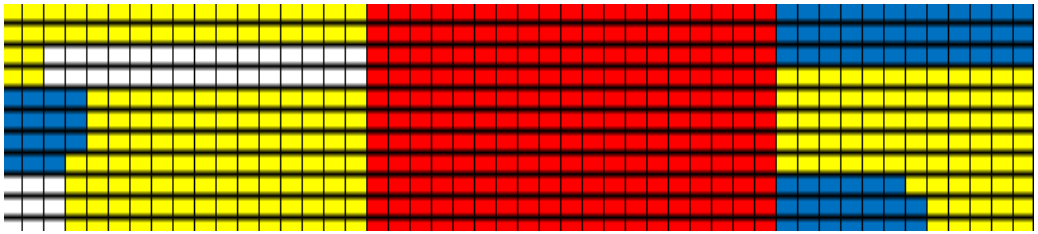
communicate compact routines affecting city-life

make patterns readily understandable and usable in applications

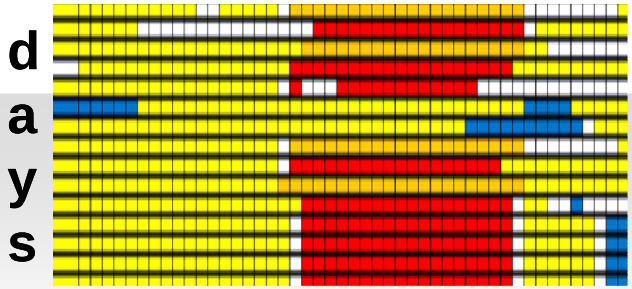
Our method

CANDIDATE LABELS POOL (E.G. "WORK 9-18", "HOME 12-14", ETC.)

LABEL PATTERN: e.g. "WORK 9-18"



REPRESENTATIONS



USER-GENERATED BEHAVIORAL PATTERNS

HHH-1	*0.1599*
HHH-2	*0.0752*
WWW-4	*0.0660*
WWW-5	*0.0372*
HHH-7	*0.0311*
EEE-5	*0.0310*
NNN-8	
HNN-8	
…	
…	

KULLBACK-LEIBLER DIVERGENCE

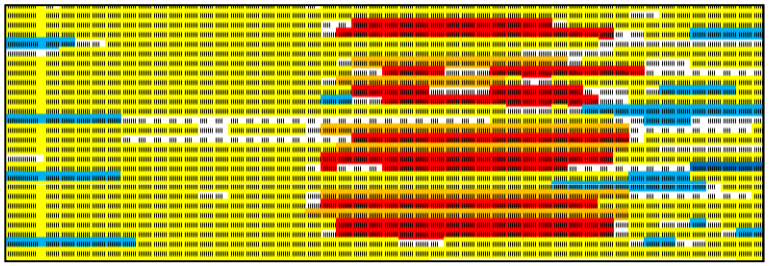
WWW-4	*0.5598*
WWW-5	*0.4978*
HHH-1	*0.0060*
NNN-2	*0.0072*
EEE-7	*0.0011*
EEE-8	*0.0010*
	0.0003
	0.0003
	0.0001

MULTINOMIAL WORD DISTRIBUTIONS

Experiments

REALITY MINING DATASET: 36 INDIVIDUALS, 121 DAYS

USER-GENERATED

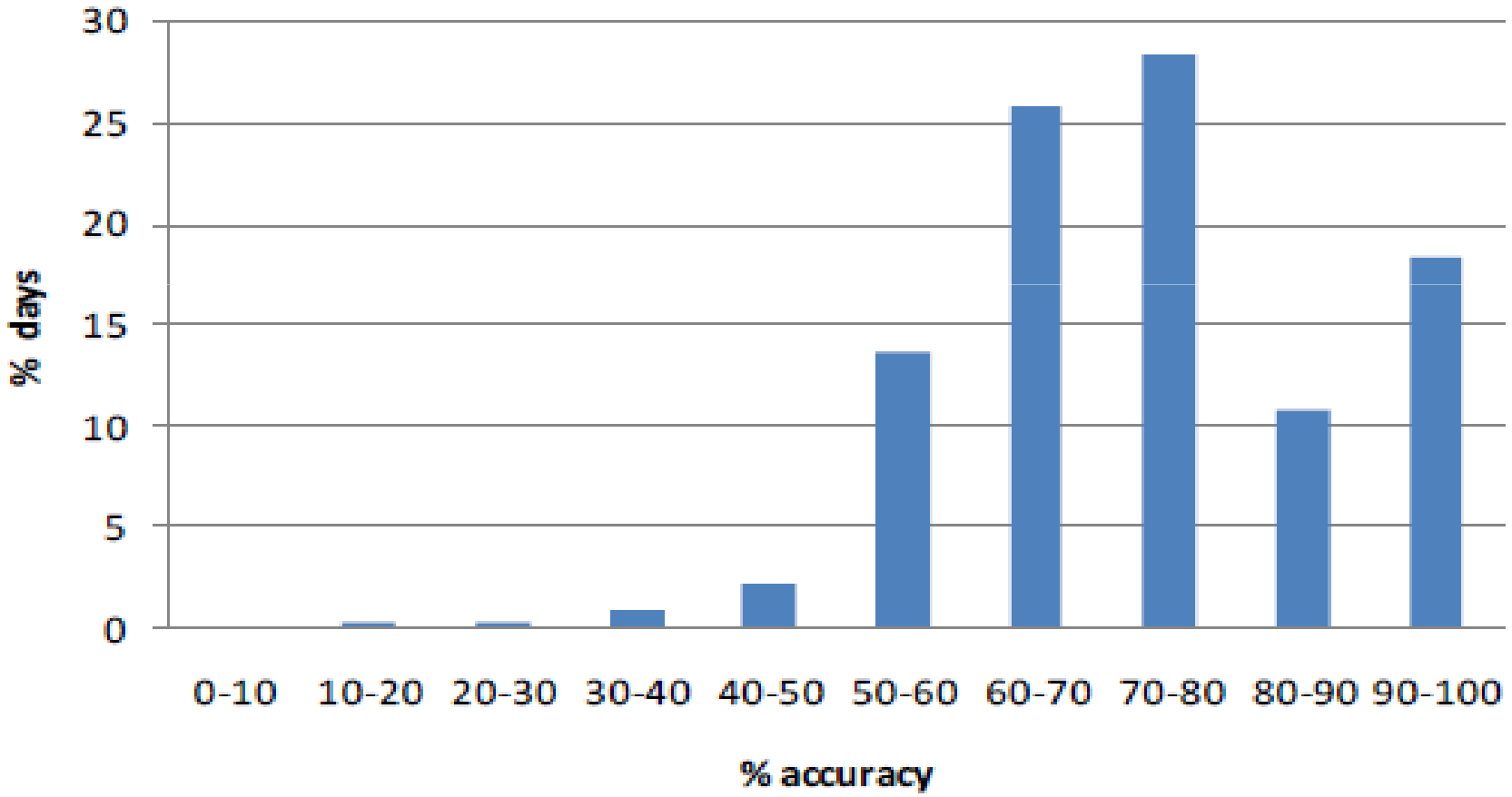


MULTINOMIAL DISTRIBUTIONS

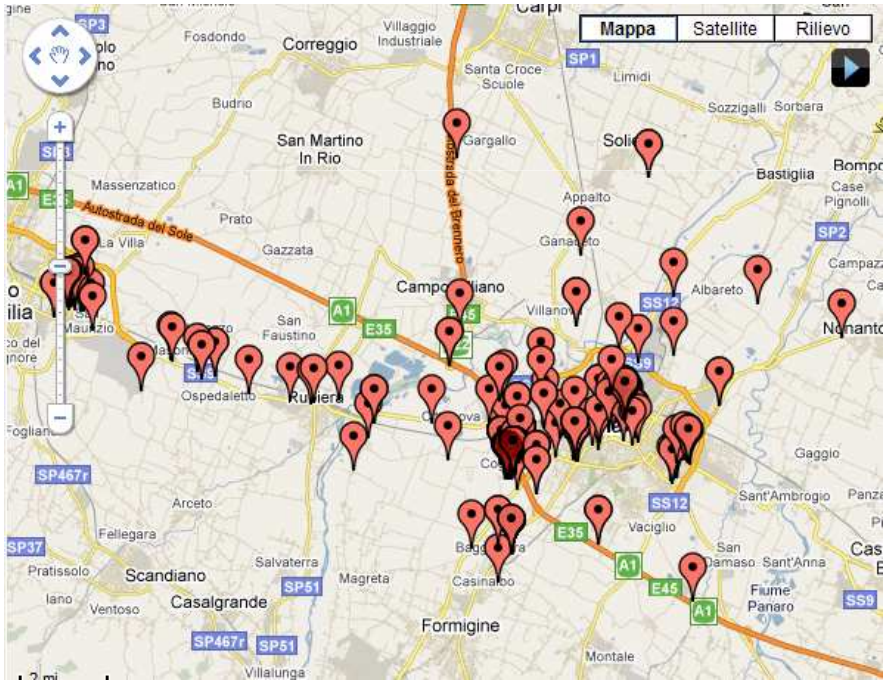
CLASSIFICATION

DAYS RECONSTRUCTION

Experiments



Google Latitude

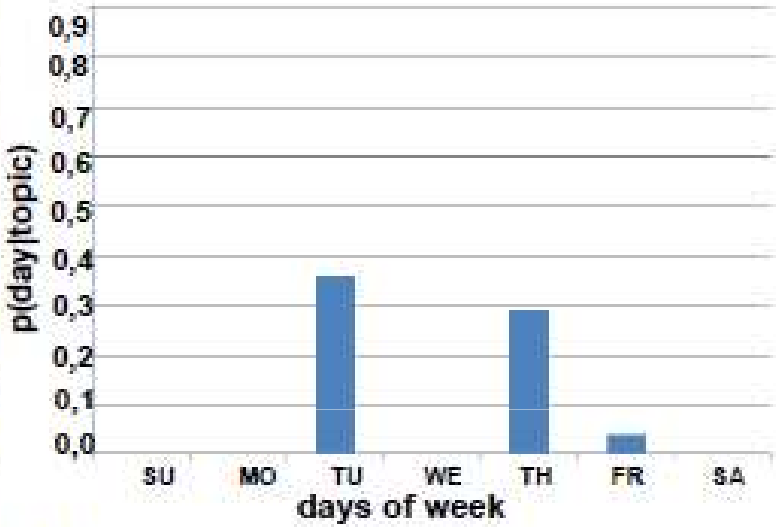


Place Discovery

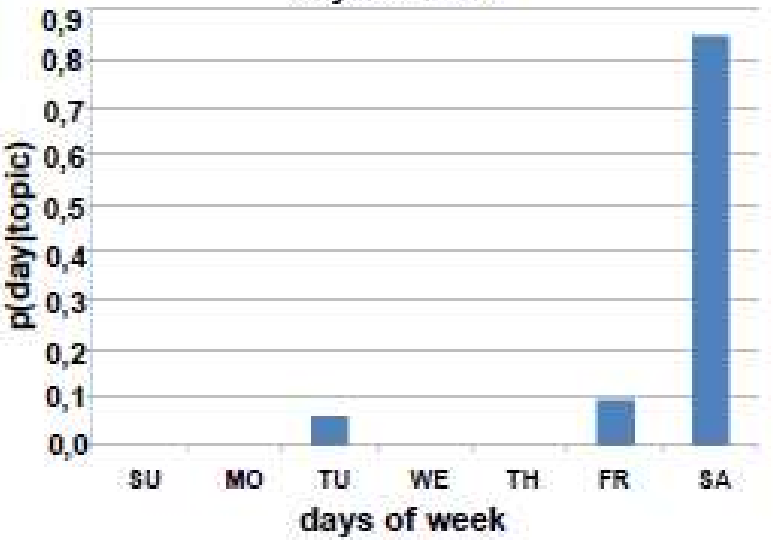
Automatic check-in!

LDA Topics


6
WWW6
WORK
4
WWW4
5
GWW5
5
WW35
GYM
HOME
3
HHH3


p(day|topic)
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
SU
MO
TU
WE
TH
FR
SA
days of week


PUB
1
PPP1
7
HNEP7
HOME
2
HHH2
3
HHH3
6
HHH6


p(day|topic)
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
SU
MO
TU
WE
TH
FR
SA
days of week

Applications

RelayRides
Neighbor-to-Neighbor Carsharing

3. Automatic Analysis of Geotagged Photos for Intelligent Tourist Services

Short-term tracking of Flickr data

Applications Scenario

- **Large database of geolocalized data** is getting available. They implicitly reveal user locations… Flickr, Twitter, Foursquares, Gwalla, Facebook Places, etc.

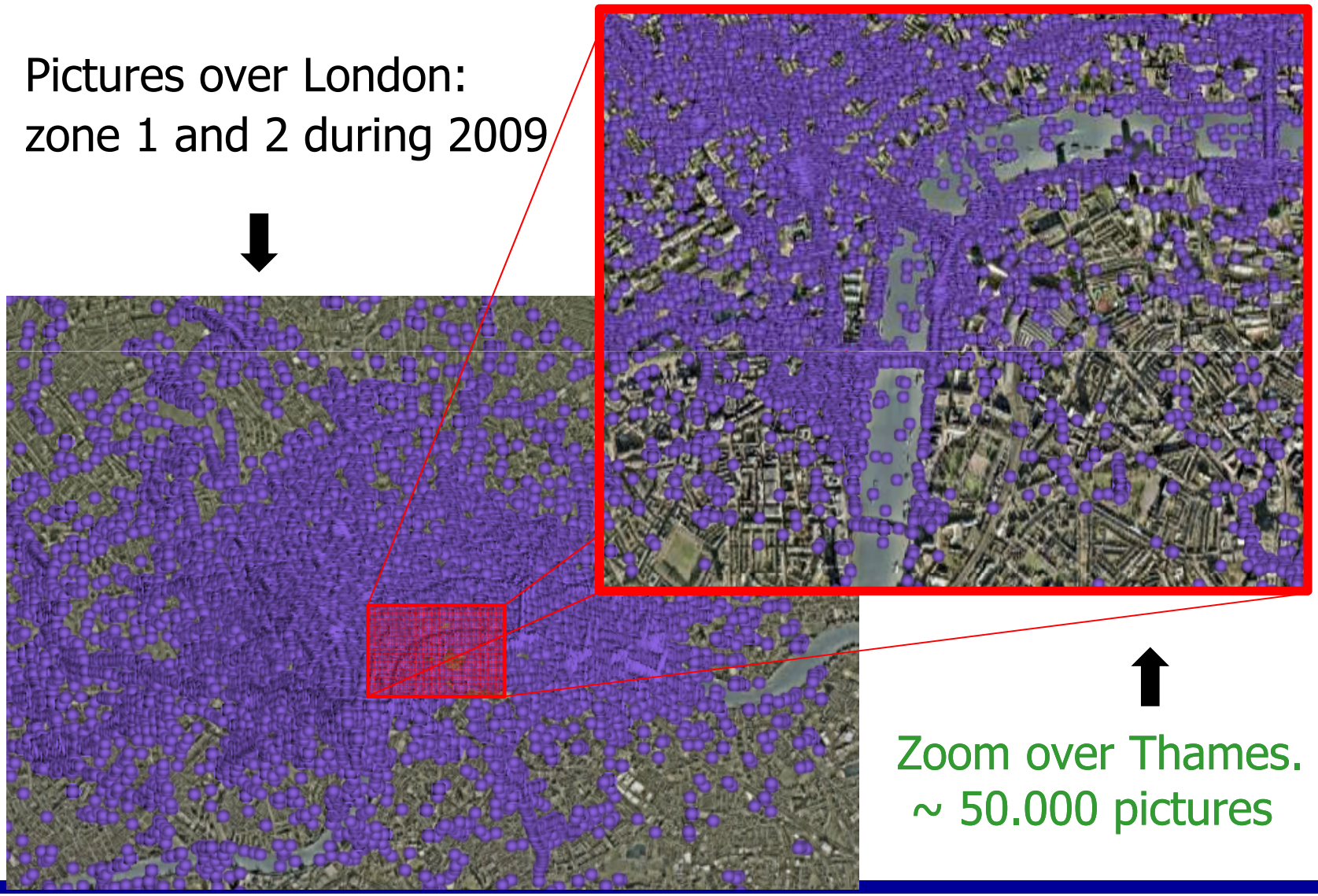
- From the extraction of such information **we foresee services to automatically aggregate and classify events**, to develop model about human/urban behaviors.

- In such context, a lot of applications and services could be developed. In particular, we concentrated in the **development of a touristic service for automatic classificatio**n and recommendations from Geotagged photos able to take advantage of **FRESH, UP to DATE, FREELY AVAILABLE** information from users.

Flicker Community

London: users upload around 180.000 pictures/year

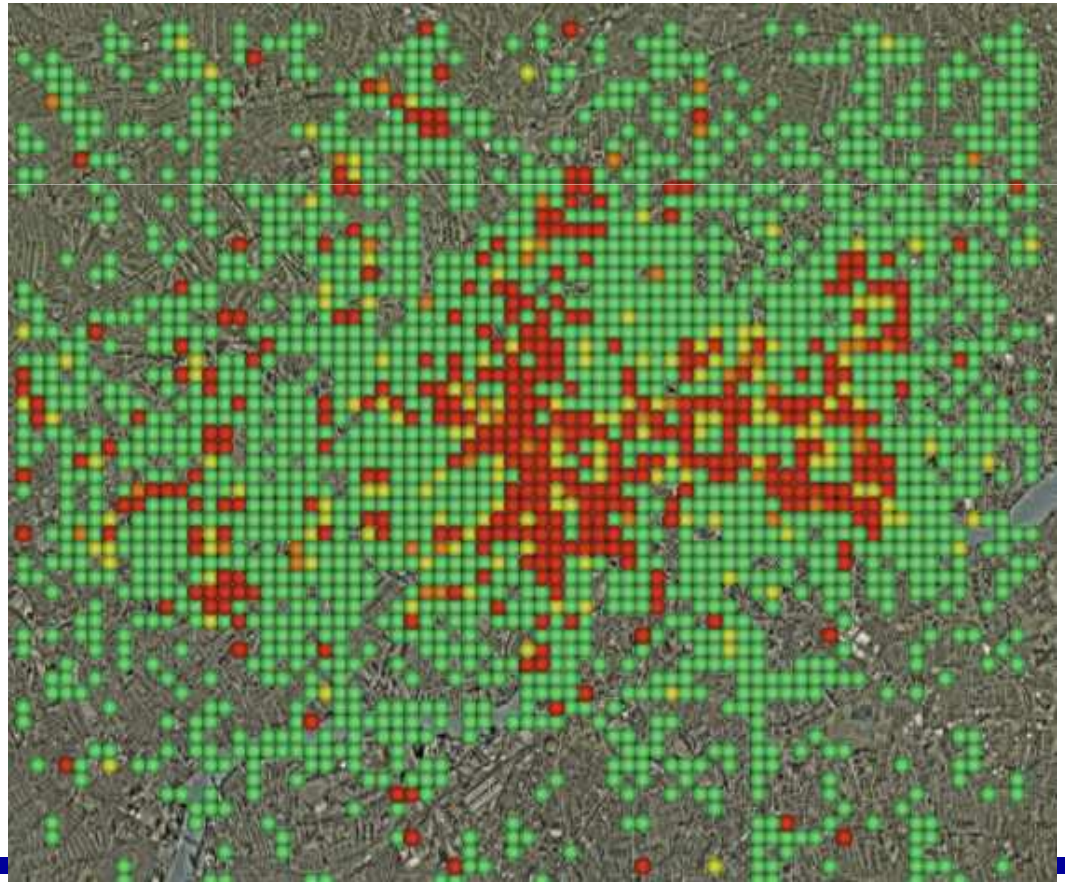
Pictures over London:
zone 1 and 2 during 2009



Zoom over Thames.
~ 50.000 pictures

Photo Clustering

- Pictures are aggregated around contiguos cells of 100x100 meters
- For each cell we count the number of pictures taken from distinct authors.
- Considering the whole number leads to noise (consider spamming user, misplaced pictures, a user taking picture to is new car, etc…)



Pictures of distinct users in a cell:

- Between 1 and 15
- Between 15 and 25
- Between 25 and 50
- Over 50

Photo Clustering (II)

- We order cells from the most “Active” one to the minor one.
- For each cell we build a **label** searching for recursive terms in picture titles or descriptions

Pictures of distinct users in a cell:

- Between 1 and 15
- Between 15 and 25
- Between 25 and 50
- Over 50

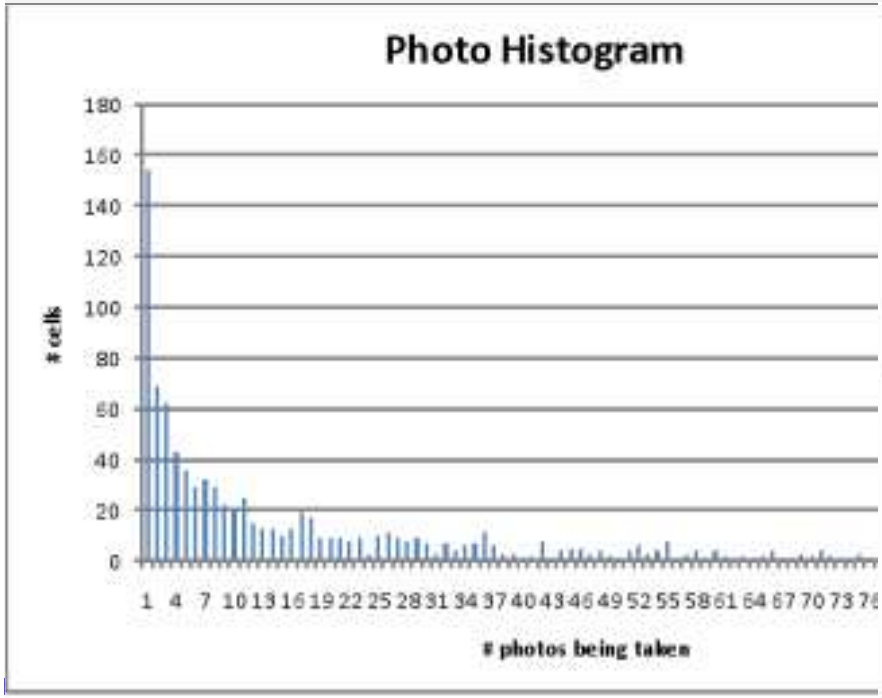
Cell selection through Otzu algorithm

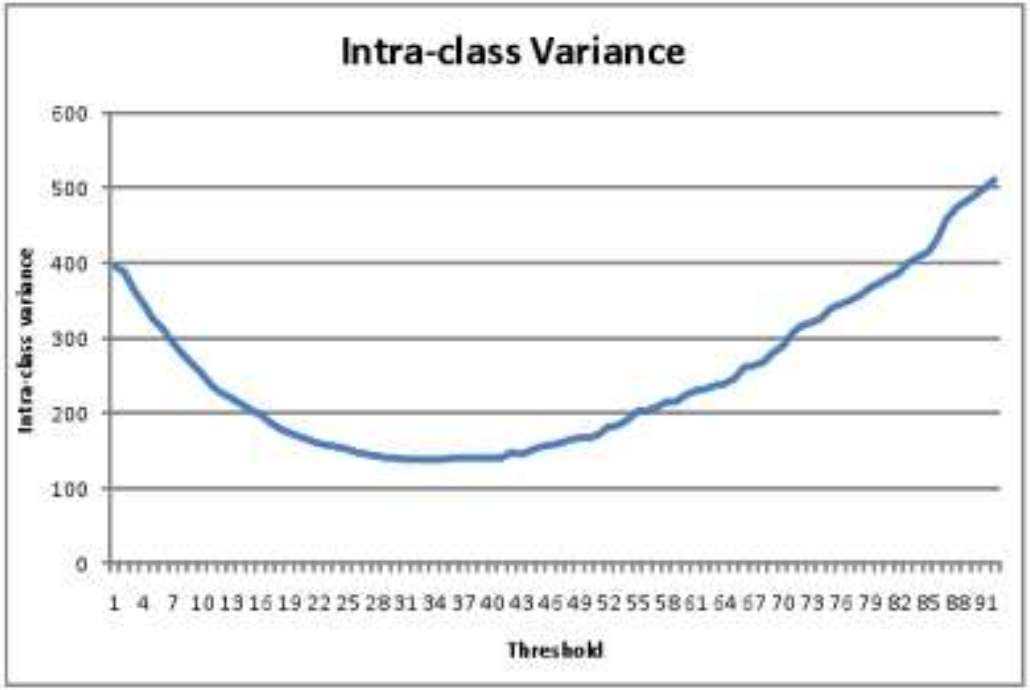
•For each possible threshold (i.e., minimum number of individual photos to mark the cell as relevant), we compute the intra-class variance between relevant and not-relevant cells (see graph on the right).

•The threshold minimizing intra-class variance is the optimal one. The algorithms consists thus in computing, for each threshold T:

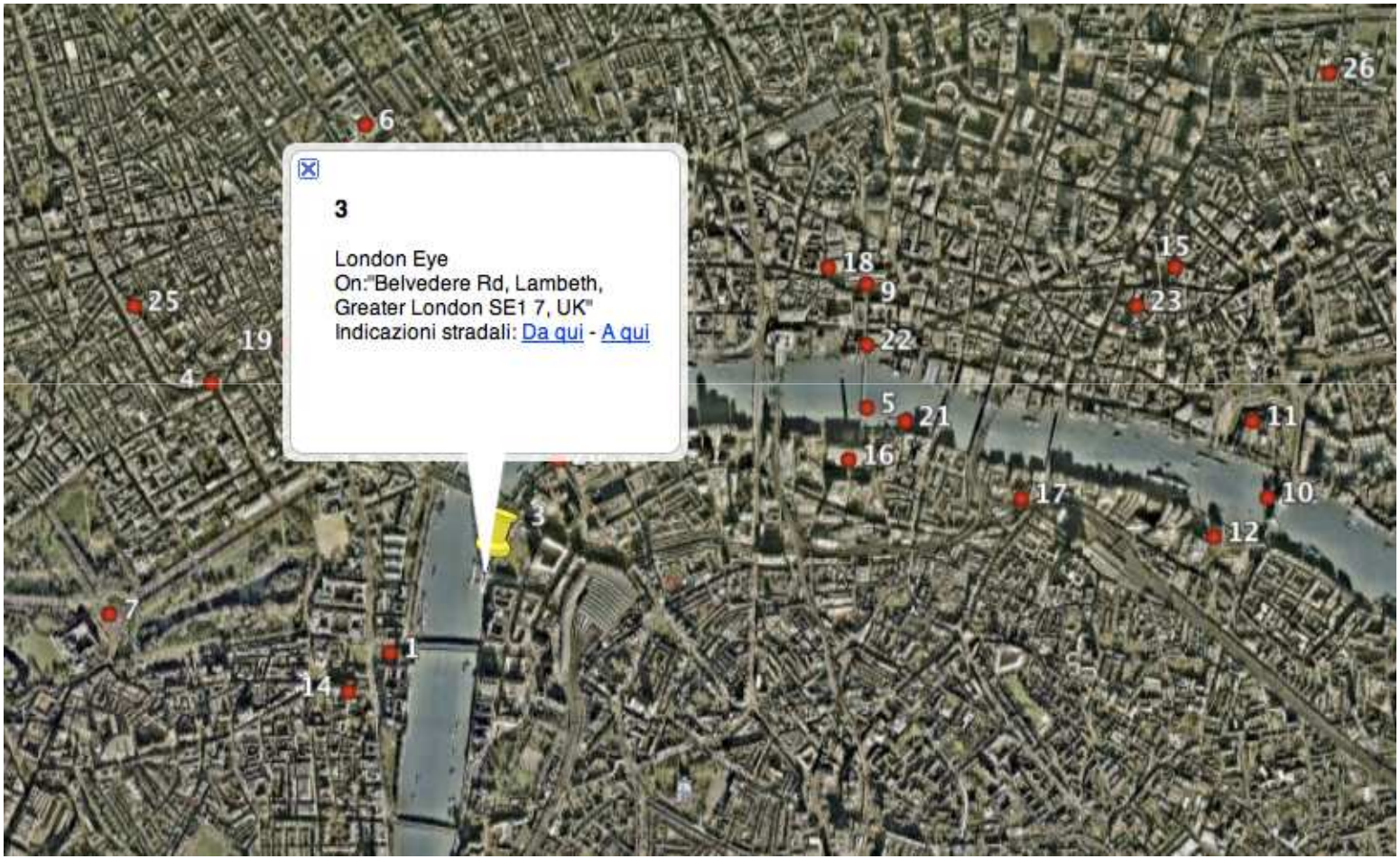
$$\sigma^2(T) = \omega_1(T) \cdot \sigma^2{}_1(T) + \omega_2(T) \cdot \sigma^2{}_2(T)$$

Where ω_1 are the probabilities of the two classes, and $\sigma^2{}_1$ are the variances of these classes





Selected cells after Otzu filtering



Results Comparison

	Berlin	London	Paris
1	Brandenburger Tor	Big Ben	Notre Dame
2	Reichstag Dome	Trafalgar Square	De Triomphe
3	Holocaust Memorial	London Eye	Sacre Coeur
4	Sony Center	Piccadilly Circus	Eiffel Tower
5	Fernsehturm Berlin	Millenium Bridge	Le Louvre
6	Berliner Dom	British Museum	Louvre Museum
7	Potsdamer Platz	Buckingham Palace	Centre Pompidou
8	Mosaik Jacket	History Museum	Dead Eyes
9	Checkpoint Charlie	Paulos Cathedral	suggestion Tour Eiffel
10	Rotes Rathaus	Tower Bridge	Terrasse des Feuillants
11	Humboldt University	suggestion The Tower	Saint Eustache
12	Neue Wache	City Hall	Place de l'Hôtel-de-Ville
13	Victory Column	Covent Garden	Palais Garnier
14	Altes Museum	Westminster Abbey	Sainte Chapelle
15	Weisse Kreuze	Mary Axe	Pont Alexandre
16	Berlin Hauptbahnhof	Tate Modern	suggestion Rue St
17	Berlin Alexanderplatz	Southwark Cathedral	Tour Eiffel
18	Christmas Market	St Paulos	suggestion Palais Royale
19	Pergamon Museum	China Town	The Pantheon
20	Eine Aktion	South Bank	suggestion Les Tuileries
21	Hackescher Markt	Globe Theatre	Looking Back
22	World Clock	Millennium Bridge	Moulin Rouge
23	Big Brother	Leadenhall Market	Petit Palais
24	Engels Forum	Camden Lock	Avenue du Général Lemonnier
25		Carnaby Street	Place Louis Lépine
26		Brick Lane	Pont Neuf
27		Leicester Square	Rainy Fountain

"Making Recomendations" based on collaborative filtering (I)

The goal is to **use the information on where a user has been before** (e.g., Franco in London**) to recommend places** he might want to visit in another city (e.g., Paris).

To perform this task**, we adopted an instance-based Pearson collaborative filtering**, also used by on-line shops (e.g., Amazon) to recommend items to users and it finds a natural application in personalized travel guides, where the attractions being proposed are tuned to the specific interests of a given user.

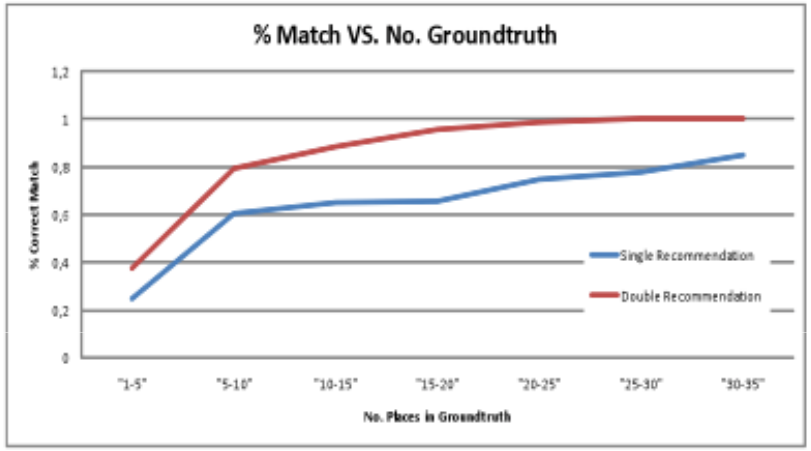
To test the performance of collaborative filtering in this scenario for each user in our dataset that visited at least two cities, we artificially removed the information on where she/he has been in a "test"-city and use the information on where she/he has been before in other cities to recommend interesting places in the "test"-city.

	London		Paris	
	Big Ben	Trafalgar Sq.	N. Dame	A. Triomphe
Alberto	yes	yes	yes	no
Marco	no	no	yes	yes
Franco	yes	yes	?	?

"Making Recomendations" based on collaborative filtering (II)

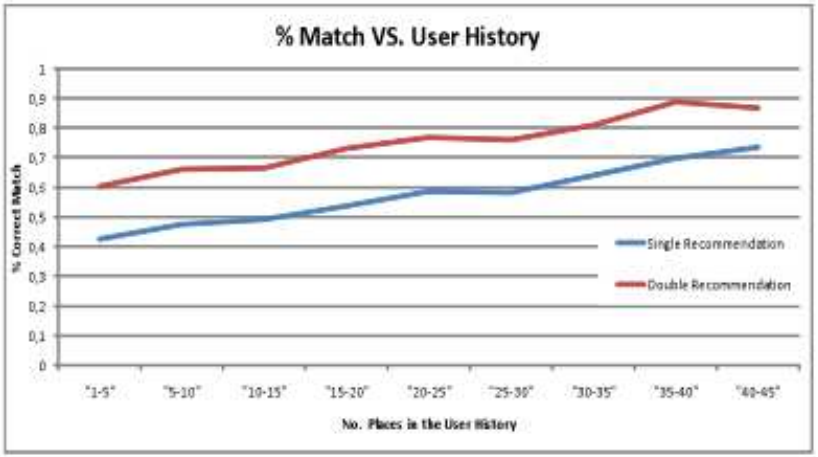
In a first set of experiments, we computed the **percent of correct recommendations on the basis of how many places the user actually visited in the test city**:

- if the **user visits only few places**, the **algorithm results not really effective in pin-pointing** (recommending) exactly those peculiar locations.
- **if the user visits a lot of places, several of our recommendations match those** places actually visited



In a second set of experiments, we performed a similar kind of analysis, but on the x-axis there **is the number of places visited *before,* by the user:**

-**more places** the user has visited before, **better recommendations** could be provided
-**good results comprise also those users to which only few spots have to be guessed**



4. Discovering large-scale city dynamics through Nokia Sports Tracker online repository of GPS tracks

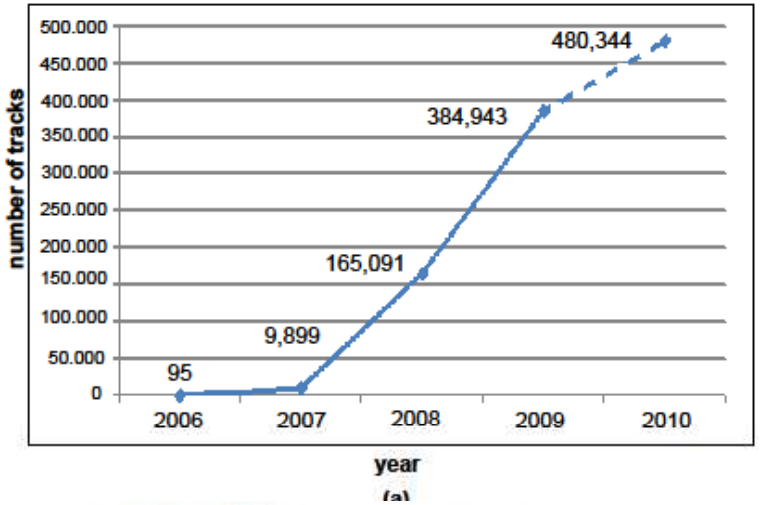
Short term on Nokia Sport Tracker Data

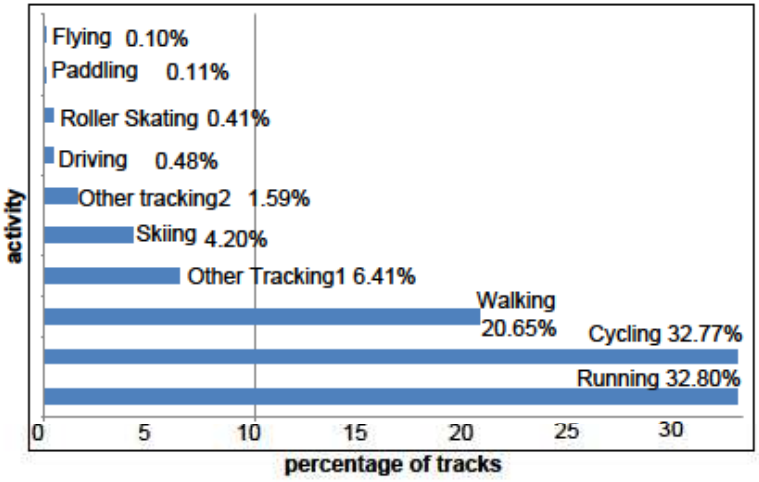
Main Idea

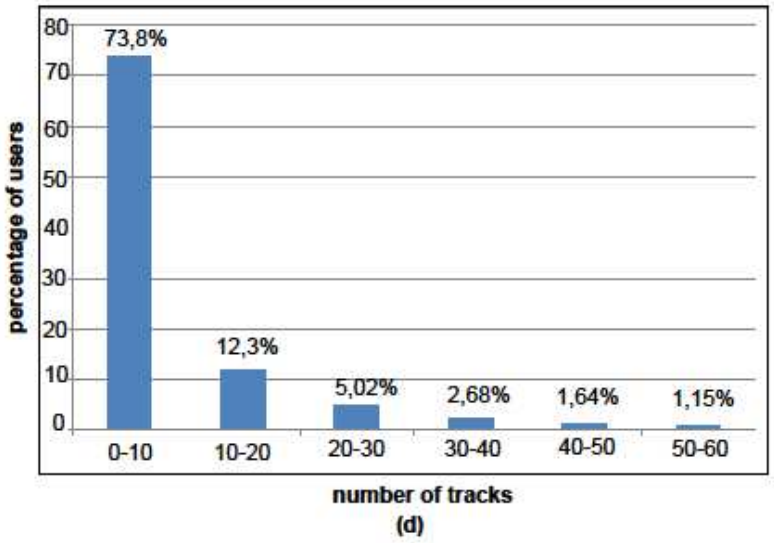
- Aggregate lots of GPS traces **annotated with the activity** the user was performing at that time to discover areas in the city where that activity is performed most.
- Also temporal analysis to discover the temporal patterns with which a given area is used.

- Nokia Sport Tracker dataset. Large (90GB) dataset of sport-annotated GPS activities.
  - Computational problems arise… need for spatial indices, and pre-computation…

Nokia Sport Tracker

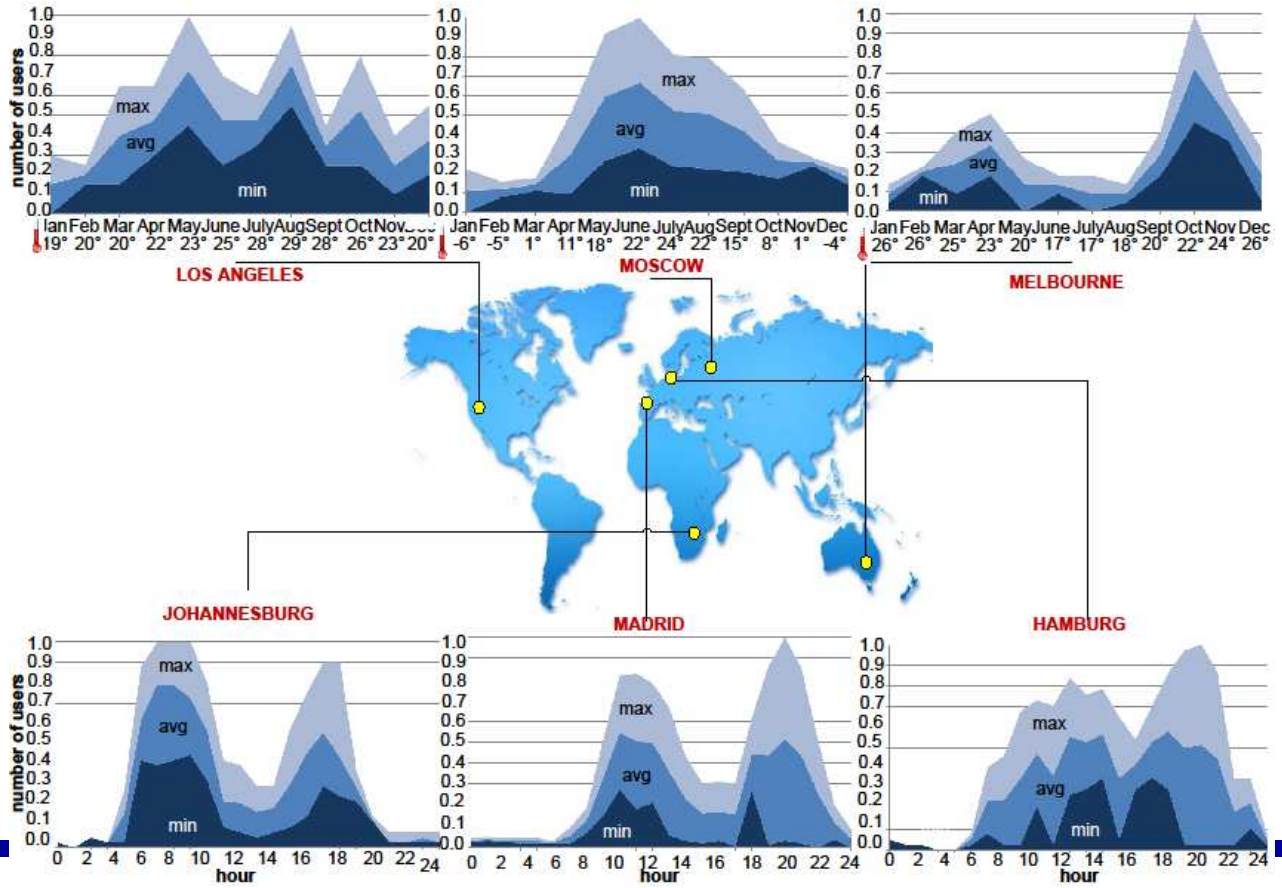






Global Temporal Analysis

- Simple statistical analyses on Nokia Sports Tracker dataset allow to highlight differences across cities.
- We computed the minimum, maximum and average of the number of users of the city on a monthly base and on an hourly base.



Finer Grain Analysis

- Apply statistical techniques to smooth individual traces in the city concerning specific activites, in order to highlight patterns and areas of interest.

- **Kernel density estimation**. is a non-parametric way of estimating the probability density function of a random variable. Given some data about a sample of a population, kernel density estimation makes it possible to extrapolate the data to the entire population

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{d(x,t_i)}{h}\right)$$

KDE Parameters

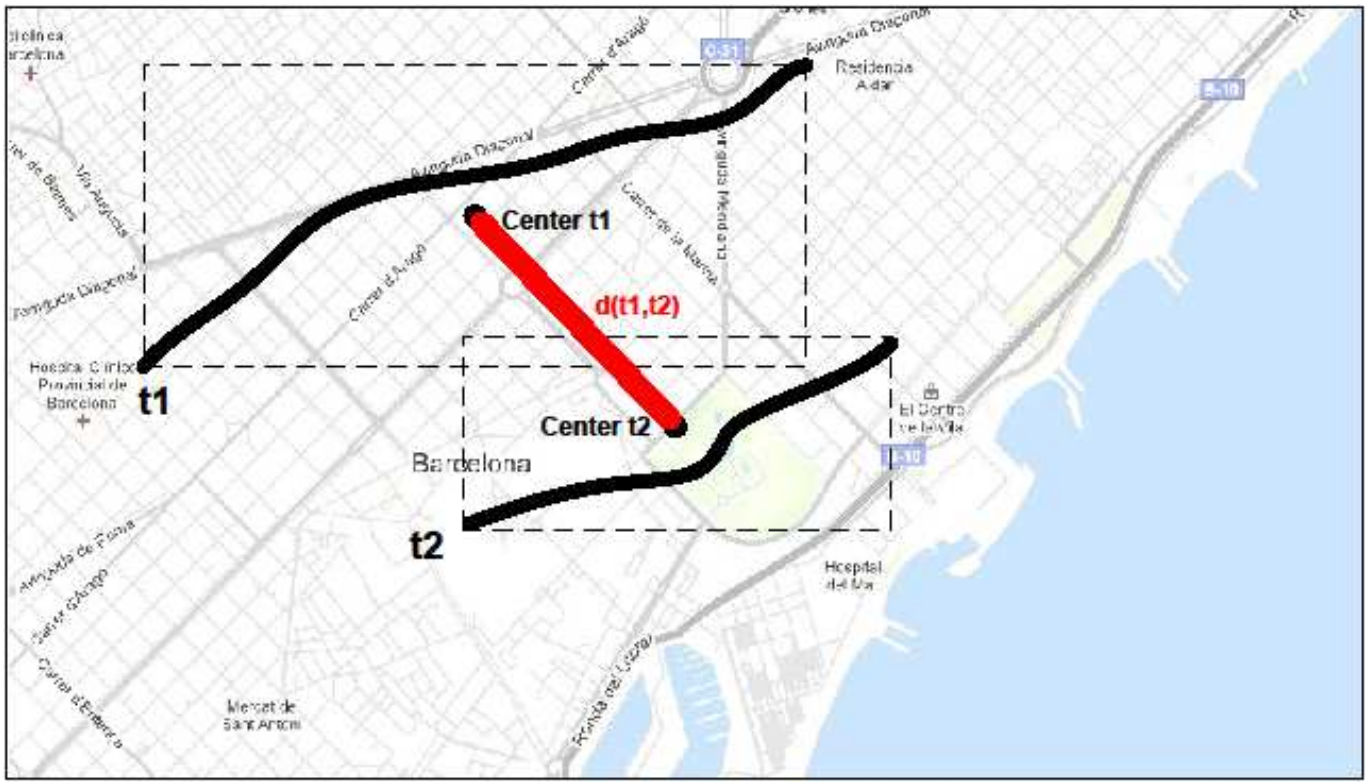
- **K** is the kenel function, it does not affect results significantly, so we used "traditional" Gaussian kernel.

$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}$$

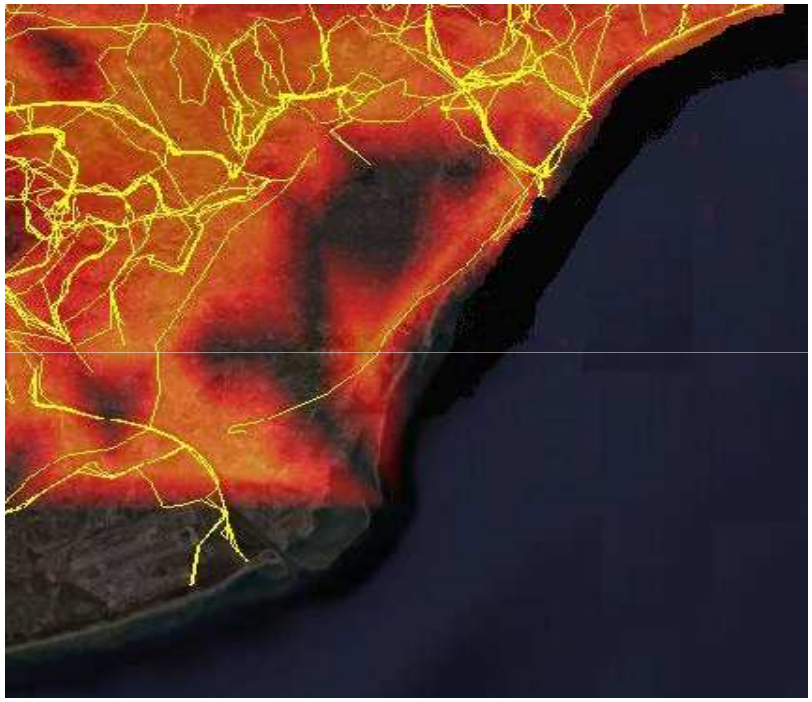
- **h** is bandwidth which controls the smoothness of the density estimate.
  - In the case of a normal distributed kernel, h represents the standard deviation of the normal distribution. The contribution of a track to the density of a point x sharply decreases as the distance from the track increases (the *68-95-99.7 rule states that for a normal* distribution, nearly all values lie within 3 standard deviations of the mean).
  - h as the average minimum separation between tracks implies that relative clusters of tracks are "collapsed" in a single peak of the density function, while the density of points farther away from all the tracks will be close to 0

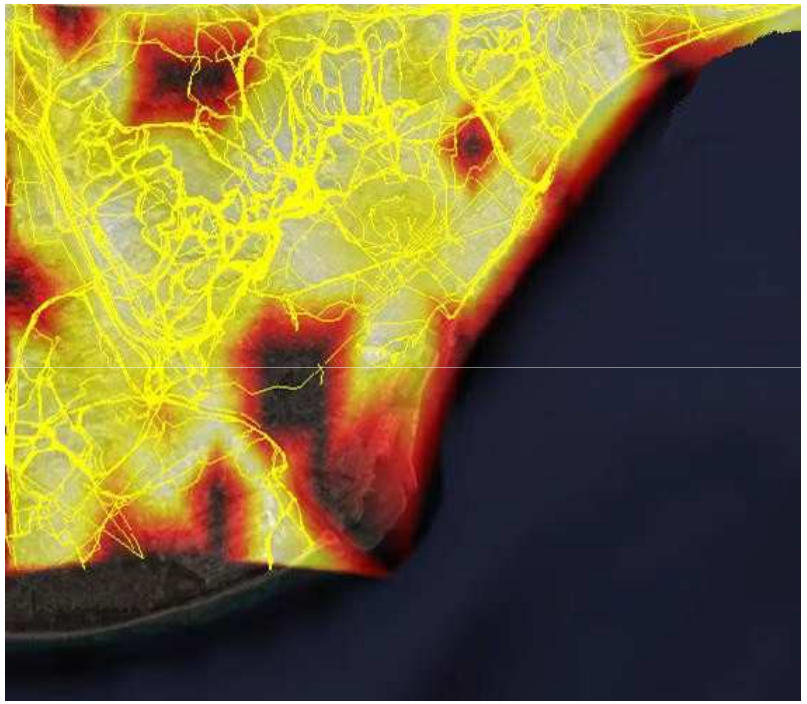
KDE Parameters

$$h = \frac{1}{N} \sum_{i=1}^{N} min\ d(t_i, t_j)\ j = 1, ..., N\ ,\ i \neq j$$



Resutls





www.mysportpals.com

Validation

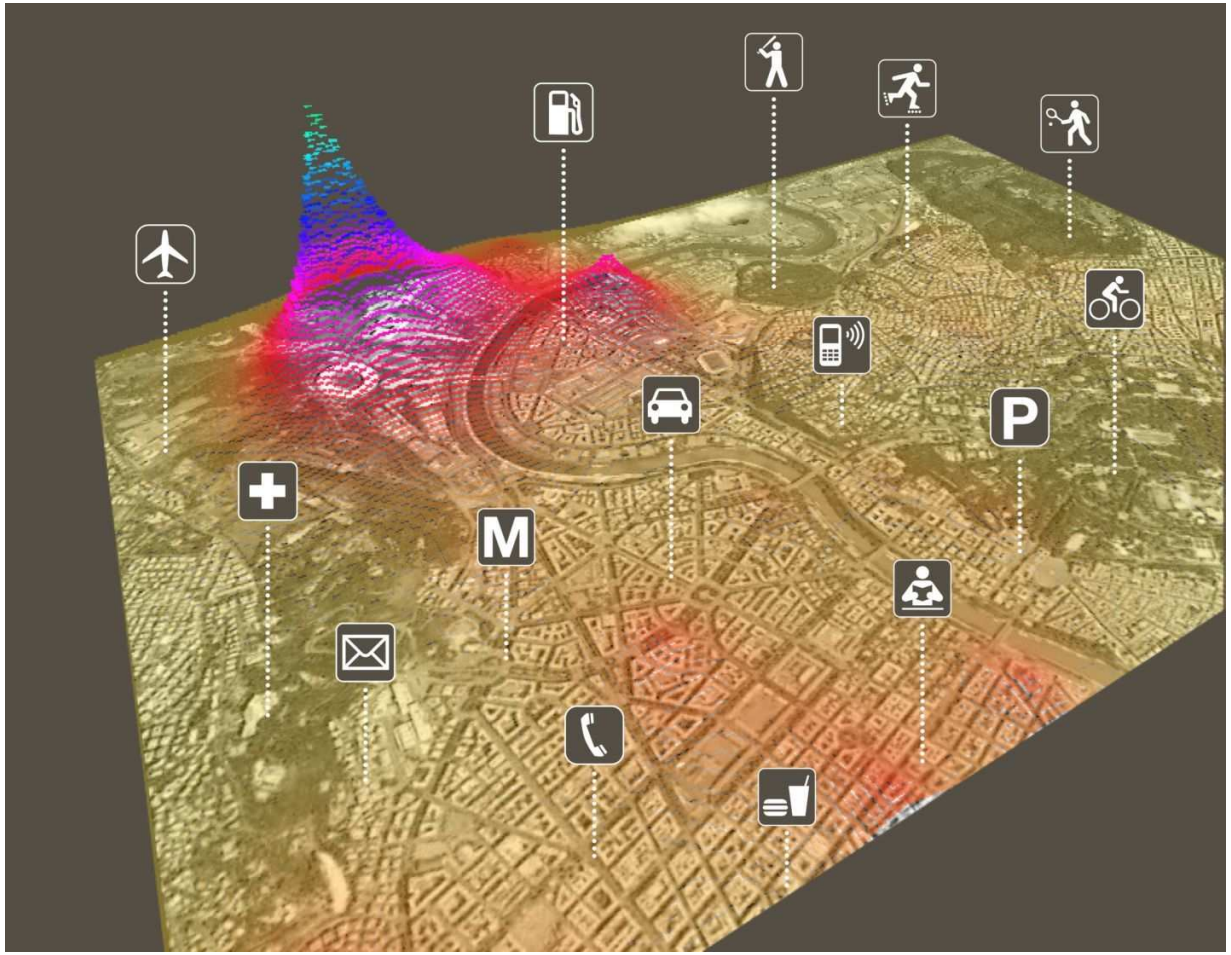
- Ok, cool… but can you validate your results?
- Difficult problem groundtruth missing…

- Compare with other dataset, looking for correlation.

- In the cycling case, we can compare obtained KDE with KDE obtained using bike routes of the city. Pearson correlation between the two distributions.

City	Bike-friendly rank	Bike-route index
Amsterdam	1	0.85
Copenhagen	2	0.63
Berlin	3	0.98
Barcelona	4	0.95

Conclusions



Future Works

- Better ways of validating results, comparison with other datasets
- Information obtained by combining different data sources
  - Mobility and yellow pages

www.mrtyp.it

- A lot of ad hoc approaches… the line between principled research and hacking becomes rather thin…
  - General approaches to analyze and visualize whereabouts data
  - General approaches to extract features from mobile data

- Techniques being developed could give hints and insights on analyzing other data (e.g., user activity on the basis of body-worn sensors)
- Life logging.